Statistical Control

Statistical control refers to the technique of separating out the effect of one particular INDEPENDENT VARIABLE from the effects of the remaining variables on the DEPENDENT VARIABLE in a MULTIVARIATE ANALYSIS. This technique is of utmost importance for making valid INFERENCES in a statistical analysis because one has to hold other variables constant in order to establish that a specific effect is due to one particular independent variable. In the statistical literature, phrases such as "holding constant," "controlling for," "accounting for," or "correcting for the influence of" are often used interchangeably to refer to the technique of statistical control. Statistical control helps us to better understand the effects of an independent variable, say, X_1, on the dependent variable Y because the influence of possibly CONFOUNDING or SUPPRESSING variables X_j ($j > 1$) is separated out by holding them constant.

How is this technically done when we estimate a particular MODEL? Suppose we find that campaign spending (X_1) increases turnout (Y) at the electoral district level. In order to establish the campaign spending effect, we want to know whether this relationship is confounded, suppressed, or SPURIOUS if we include control variables in our model that account for alternative explanations. The expected closeness of a district race (X_2) is such an alternative explanation. We expect that close races should particularly motivate voters to turn out on Election Day. Consider the following LINEAR regression model,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon.$$

The specific effect β_1 of campaign spending on turnout in the classical regression framework is estimated as

$$\hat{\beta}_1 = \frac{\sum(X_1 - \hat{X}_1)(Y - \hat{Y})}{\sum(X_1 - \hat{X}_1)^2},$$

where $\hat{X}_1 = \alpha_0 + \alpha_1 X_2$ is predicted from a regression of X_1 on X_2; and $\hat{Y} = \gamma_1 + \gamma_2 X_2$ is predicted from a regression of Y on X_2. Thus, the estimated campaign spending effect $\hat{\beta}_1$ solely depends on variation in X_1 and Y, because any linear impact of X_2 in the terms $X_1 - \hat{X}_1$ and $Y - \hat{Y}$ is subtracted. Consequently, $\hat{\beta}_1$ is calculated based on terms that are independent of X_2. When estimating the specific effect of X_1 on Y, we hold constant any possible distortion due to X_2 because the relationship between X_1 and Y no longer depends on X_2. Thus, statistical control separates out the specific effect $\hat{\beta}_1$ of X_1 on Y corrected for any influence of X_2. At the same time, statistical control also disentangles the specific effect $\hat{\beta}_2$ of X_2 on Y corrected for any influence of X_1. Thus, this technique does not treat our key explanatory variable any different from other (control) variables. The logic behind statistical control can be extended to more independent variables even for GENERALIZED LINEAR MODELS.

Statistical control comes with simplifying ASSUMPTIONS, though. Ideally, one likes to implicitly hold confounding factors constant, such as in experimental control through randomization. Because this is often not possible with observational data, one has to explicitly identify factors before one can statistically control for them. Furthermore, statistical control assumes that possible confounding effects are linear and

additive, that is, they are independent of the particular level of any other variable in the model. Coming back to the running example, the expected closeness of the race is assumed to have the same impact on campaign spending and turnout no matter how close the race is expected to be. If these assumptions prove to be wrong, one can include product terms of the variables in the model to control for possible CONDITIONAL or INTERACTION EFFECTS or restrict estimation of the model to subpopulations in which the assumptions hold.

Thomas Gschwend

References

Gujarati, D. N. (2003). *Basic econometrics* (4th ed.). Boston: McGraw-Hill.

Lewis-Beck, M. S. (1995). *Data analysis: An introduction.* Thousand Oaks, CA: Sage.

Mosteller, F., & Tukey, J. W. (1977). *Data analysis and regression*. Reading, MA: Addison-Wesley.