Yearbook of Population Research in Finland 33 (1996), pp. 328–332

A Note on the Use of Anticipatory Covariates in Event History Analysis

JUHA M. ALHO

Professor of Statistics
Department of Statistics
University of Joensuu
Joensuu, Finland

Abstract

Anticipatory covariates are regressors whose values become known only after the value of the dependent variable has been ascertained. Hoem (1995) has given an informal discussion concerning the possible pitfalls in the use of such covariates in event history analysis. This paper complements Hoem's findings by using simple linear regression as the framework. It turns out that complex patterns of bias may be introduced by the use of anticipatory covariates. In all cases it may not be possible to guarantee that the magnitude of the bias remains small. Therefore, extreme care is needed in interpreting results from studies that have relied on anticipatory covariates.

Keywords: anticipatory covariates, event history analysis

Introduction

Hoem (1995) gives a careful discussion of the dangers of anticipatory covariates in event history analysis. These are regressors whose values are determined only after the value of the dependent variable has been ascertained. In Hoem's example education is used to explain divorce risk, but the level of education is known only at the end of the follow-up period. Hence, it is called "anticipatory". The main result is that if the occurrence of divorce does not have a strong influence on the subsequent educational history, then the use of the anticipatory education measure, instead of the actual level of education at the time of (or before) marriage, tends to "dampen" the estimated association.

Hoem gives an informal but convincing argument to support his conclusions:

> "... assume for instance that there is a negative gradient in the educational effect on divorce risks. Suppose that we select the lowest level of education as our basis of comparison... Women who have this level on our anticipatory variable have also had this level throughout, so their divorce risk should be correctly assessed by the practice that we follow. Women at higher final level of education must have been subject to the higher divorce risks that pertain to lower levels of education during part of their life, however, so our assessment of their divorce risk should be too high, as compared to what we would get if we had known their educational level at all times during their adult lives."

Especially valuable are his empirical calculations for the Swedish cohorts of 1959 and 1964 for which both the true and anticipatory covariates are available. Hoem's Table 1 shows that for those marrying at the age of 16–17, and to a lesser extent those marrying at the age of 18–19, the effect of education on the risk of divorce is severely overestimated by the use of anticipatory education measure as compared to the actual level of education at the time of divorce. The explanation seems to be that for those marrying young, divorce opens up a new chance to continue their education (Hoem 1995).

Although a formal discussion of the problem would be interesting, a related paper of Gail (1986) demonstrates that a discussion of general relative risk regression is difficult and leads to complex mathematics. The purpose of this note is to complement Hoem (1995) at a more elementary level. We use some well-known aspects of ordinary least squares regression theory to discuss the use of anticipatory covariates. The level of presentation is similar to that of Hoem (1992). It turns out that there are other possible pitfalls besides those mentioned in Hoem (1995).

Single covariate

Suppose Y is the dependent variable, such as some measure of divorce risk, and let X be an explanatory variable, such as education at the time of marriage. Assume that the true model is

$$Y = a + bX + \varepsilon, \tag{1}$$

where a and b are coefficients, and ε is an error term (with mean zero and independent of X), so that the conditional expectation of Y given X is of the form $E[Y \mid X] = a + bX$. Suppose Z measures the same property as X (say, education), but at a later time, when Y (the possible occurrence of divorce) has already become known. Assume that

$$Z = c + dY + eX + \eta, \tag{2}$$

where c, d and e are coefficients, and η is an error term independent of X, Y and ε.

Often we would have $e = 1$ in (2). For example, suppose Y is centered to have a mean of zero. Then, later education Z would be equal to the earlier education X plus the average increment c, as modified by the possible positive or negative effect of Y and by residual random variation η. Or, the increment would be $Z - X = c + dY + \eta$, and therefore, $Z = c + dY + X + \eta$. However, if earlier education would also influence the increment with a coefficient e', then we would have $Z - X = c + dY + e'X + \eta$, and $Z = c + dY + eX + \eta$, where $e = 1 + e'$. It will be useful to keep the possibility of $e \neq 1$ open.

Suppose that we do not know the value of X, so instead of (1) we fit the model

$$E[Y \mid Z] = \alpha + \beta Z.$$

If the first two moments of Y and Z are known, the least squares estimate of the slope is

$$\beta = \frac{\mathrm{Cov}(Y, Z)}{\mathrm{Var}(Z)}.$$

(Here, and in the sequel, we will phrase our arguments in terms of the true population moments. Under mild regularity conditions the actual empirical least squares estima-

tors can be shown to converge to these quantities, when sample size approaches infinity.) Under (1) and (2) we have

$$\begin{aligned}\mathrm{Cov}(Y, Z) &= \mathrm{Cov}(a + bX + \varepsilon,\ c + d(a + bX + \varepsilon) + eX + \eta) \\ &= b(db + e)\mathrm{Var}(X) + d\mathrm{Var}(\varepsilon),\end{aligned}$$

and

$$\mathrm{Var}(Z) = (db + e)^2\mathrm{Var}(X) + d^2\mathrm{Var}(\varepsilon) + \mathrm{Var}(\eta).$$

Therefore, we have the general representation

$$\beta = b\ \{\frac{(db + e)\mathrm{Var}(X) + d\mathrm{Var}(\varepsilon)/b}{(db + e)^2\mathrm{Var}(X) + d^2\mathrm{Var}(\varepsilon) + \mathrm{Var}(\eta)}\}.$$

By taking $d = 0$ in (3), i.e., by assuming that the dependent variable does not influence the evolution of X over time, we get

$$\beta = b/e\ \{\frac{e^2\mathrm{Var}(X)}{e^2\mathrm{Var}(X) + \mathrm{Var}(\eta)}\}.$$

Requiring further that $e = 1$ yields a form of the principal result of Hoem (1995), or

$$\beta = b\ \{\frac{\mathrm{Var}(X)}{\mathrm{Var}(X) + \mathrm{Var}(\eta)}\}, \qquad (4)$$

so β has the same sign as b, but $|\beta| \leq |b|$. In other words, the use of an anticipatory regressor in place of the correct one will bias its coefficient towards zero.

Note however, that (4) requires not only that $d = 0$, but also that $e = 1$. For example, if past education had a negative effect on the subsequent increment of education (or $e < 1$) and $\mathrm{Var}(\eta)$ were small, then the use of an anticipatory education measure might even lead to an overestimation of the effect of education on divorce risk.

A reader familiar with measurement error models will recognize (4) as the classical result concerning regression coefficients when explanatory variables have been measured unbiasedly, but with some random error (Fuller 1987, 3). In our case η corresponds to measurement error and causes the bias.

Hoem (1995) points out that if divorce influences later opportunities of education, then the use of anticipatory covariates may ruin the whole regression. This can be seen in our framework. Take for example $e = 1$ and suppose $d = -1/b$ (the exact equalities are not important here; we use them just to get a simple formula). Then,

$$\beta = -b\ \{\frac{\mathrm{Var}(\varepsilon)}{\mathrm{Var}(\varepsilon) + b^2\mathrm{Var}(\eta)}\},$$

so even if we would have $\mathrm{Var}(\eta) = 0$, the least squares estimate $\beta = -b$ would be quite different from b!

Several covariates

The regression formulation can be used to derive many other properties of the anticipatory covariate modeling. When several covariates are present in regression, it is possible that the use of anticipatory versions for some of them may bias the estimates of the others. The intuitive reason for this is that the anticipatory covariates may "eat up" some of the effect of the other covariates.

Consider, for example, the case in which divorce is influenced by both education X_1 and birth year X_2, so

$$Y = a + b_1X_1 + b_2X_2 + \varepsilon. \tag{5}$$

Suppose that the later increment in education is similarly influenced by both:

$$Z = c + dY + e_1X_1 + e_2X_2 + \eta.$$

Again, we would often – but not always – have $e_1 = 1$.

Suppose we don't have X_1 available and we use least squares to fit the model

$$E[Y \mid Z, X_2] = \alpha + \beta_1 Z + \beta_2 X_2.$$

If the first two moments of Y, Z and X_2 are known, an analogue of (3) can be given, but it is more complicated because it involves both β_1 and β_2. We omit those details, but present a simpler discussion of a special case, based on the representation

$$Z = c + da + (db_1 + e_1)X_1 + (db_2 + e_2)X_2 + d\varepsilon + \eta.$$

Let us consider the simplest case of $d = 0$, so $Z = c + e_1X_1 + e_2X_2 + \eta$. Note first that the role of η is the same as before: it biases estimates towards zero. Let us simplify further, and take $\eta = 0$ also. In this case Z and X_2 span the same space as X_1 and X_2, so the use of Z and X_2 produces exactly the same fit as the use of X_1 and X_2. We can now write the conditional expectation of Y given X_1 and X_2 in terms of the new parameters α, β_1, and β_2, as follows

$$E[Y \mid X_1, X_2] = \alpha + \beta_1 c + \beta_1 e_1 X_1 + (\beta_1 e_2 + \beta_2)X_2. \tag{6}$$

Using model (5) we can also write

$$E[Y \mid X_1, X_2] = a + b_1X_1 + b_2X_2. \tag{7}$$

The coefficients of X_1 and X_2 must be equal under both representations (6) and (7), so we must have

$$\beta_1 = b_1/e_1;\ \beta_2 = b_2 - b_1e_2/e_1.$$

We see that even if $e_1 = 1$, the coefficient β_2 will remain biased for b_2.

For example, if later birth cohorts increase their educational level more than earlier cohorts (or $e_2 > 0$), and, say, education decreases divorce risk ($b_1 < 0$), then the effect of birth cohort on divorce risk (or b_2) will be overestimated, if anticipatory Z is used in place of X_1. The degree of bias depends on e_2. The intuitive cause for the bias is that the later education Z measures not only the earlier education X_1, but also the birth cohort effect X_2.

Discussion

Above, we have tried to complement the results of Hoem (1995) by simple regression arguments. Slight modifications may be necessary when applying the results to other relative risk models. We suspect the situation may be analogous to that studied by Gail (1986). He has shown, for example, that omitting a "balanced" covariate from regression (i.e., a covariate that has the same distribution among the exposed and the nonexposed in an epidemiological study) does not cause bias in the kind of linear model we have considered, but it does cause bias towards zero in Cox regression (Gail 1986, 6). A similar effect might well exist in our setting.

The motive for using anticipatory covariates in relative risk regression is that proper data may not be available. In some cases a researcher may be able to argue that, say, the level of education cannot have changed during the follow-up period because the participants had passed the typical formative years. In this case the use of anticipatory data may be quite reasonable. However, in view of the possibility of very complex patterns of bias caused by anticipatory covariates, it may often be safer to conclude that the data do not permit the intended study than to try to force through an analysis whose validity cannot be guaranteed.

Above we have illustrated just a few of the possible biases. In particular, no correlations between ε and η were considered. The problems are compounded, if we admit the possibility that some important covariates have not been measured at all.

References

Fuller, W.A. 1987. Measurement Error Models. New York: Wiley.

Gail, M.H. 1986. Adjusting for covariates that have the same distribution in exposed and unexposed cohorts. In: Modern Statistical Methods in Chronic Disease Epidemiology, edited by S.H. Moolgavkar and R.L. Prentice, pp. 3–18. New York: Wiley.

Hoem, J.M. 1992. Harmless model misspecification. In: Mennesker og modeller, edited by O. Ljones, B. Moen and L. Østby, pp. 135–145. Oslo: Statistisk Sentralbyrå.

Hoem, J.M. 1995. The harmfulness or harmlessness of using an anticipatory regressor: how dangerous is it to use education achieved as of 1990 in the analysis of divorce risks in earlier years? Paper presented at the 11th Nordic Demographic Symposium, Helsinki, June 11–13, 1995. Published in this yearbook, pp. 34–43.