Period and cohort “effects”: interesting history, weak science

Imagine two samples of people who were 50 years old at the time of sampling: one sample was drawn in 2000 and the other – ten years later, in 2010. Someone collected information on smoking habits, computed the percentage of smokers in each sample, and reported a decline: from 30% in 2000 to 20% in 2010. What have we learned from this study?

A simple-minded answer will restate the results: we learned that the prevalence of smoking has declined from 30% to 20%. We didn’t know what happened in those years, and now we know. But what kind of knowledge have we gained? Is it historical knowledge, similar to learning about the number of Democrats in the U.S. Senate in 2010, or finding out the percentage of countries that voted for a U.N. resolution?

Most epidemiologists consider themselves scientists, not historians, and they will claim to have discovered something about cause and effect. They will name three variables: *AGE* – fixed here to 50 by design; *PERIOD* – the sampling year (2000, 2010); and *COHORT* – the birth year of the participants (1950, 1960). Then, they will talk about estimating the age-period-cohort effects on smoking, often abbreviated as APC analysis.

Two intractable problems, however, are built into APC analysis: one is scientific and the other is technical.

The scientific problem

First and foremost, there is no such thing as the effect of the calendar year (birth year, sampling year), or the effect of subtracting one calendar year from another (age). Time is the medium in which causation operates, but it is not a cause of anything. Although *AGE*, *PERIOD*, and *COHORT* are variables in a statistical sense (they take values), they do not belong to the set of natural variables that make the fabric of causal reality. Not being intrinsic properties of objects, they have neither causes nor effects. Anyone who thinks differently should read and think about thought bias.[1,2]

AGE, *PERIOD*, and *COHORT* are *associated* with various outcomes because they substitute for some natural, causal variables. For instance, *AGE* may substitute for properties such as cognitive function and DNA repair, which tend to deteriorate as a person ages. *PERIOD* may substitute for properties such as prescribed drugs and air pollution, which may differ between survey years. And *COHORT* may substitute for causal variables whose values depend on the era during which one lived: anything between the beginning of pregnancy and just before the value of *PERIOD*.

Many epidemiologists will probably agree with what I wrote so far. They will claim, however, that this is precisely what is estimated by APC analysis: the effect of all kinds of natural variables that are captured by *AGE*, *PERIOD*, and *COHORT*. If I specify the outcome, they might even be willing to start a list of candidate variables.

And there lies the scientific problem.

A causal theory is not an open list of “likely” and “possible” and “unknown” variables. It is a bold claim about causal reality that runs the risk of being wrong, and may be immediately challenged. When someone claims, for example, to have estimated the effect of mammography on survival, an army of critics is on guard. Was the variance large or small? Was confounding by variable *Z* accounted for? Was colliding bias added by mistake? What about information bias? Effect modification bias? Causal pathway bias?[2]

APC theories are immune to many of these challenging questions, because the variables *AGE*, *PERIOD*, and *COHORT* do not claim much about causal reality. Typical APC analysis does not follow a causal diagram, so there are no external confounders, colliders, or modifiers to consider.[2] Whichever estimate is computed, it is not an estimate of any causal parameter. In the opening example, for instance, the prevalence difference in smoking – the so-called *PERIOD* effect – arose from an unknown combination of parameter estimates for some list of causal variables. Which variables and which estimates make up the list? Don’t ask. We plead the Fifth.

Since APC theories do not claim much about causal reality, they do not expose themselves to many challenges. *But they do not add much knowledge, either*. It was Karl Popper who astutely pointed out the relation between the content of a theory (knowledge) and its susceptibility to challenges (falsifiability).[3] The greater the informative content of a theory, the more susceptible it is to empirical challenges, and the greater the knowledge we may gain if the theory survives. Conversely, a theory that does not face many challenges must be thin in content; it does not add much knowledge.

What challenges are there for a local theory about the prevalence of smoking at two times, which does not even claim to estimate a causal parameter? At most, we may raise questions about the sample size, the sampling, and the measurement. That's about it. Compare these limited challenges, for example, to the ongoing controversy about the effect of mammography on survival.

The technical problem

The technical problem of APC analysis was recognized long ago[4] (although it is often ignored in epidemiology textbooks). To explain the matter, I will assume for a moment that *AGE*, *PERIOD*, and *COHORT* are causal variables. Is it possible to estimate their effects on smoking?

The difficulty is already apparent in the opening example. Can we tell which effect is estimated by the prevalence difference in smoking between 2000 and 2010? It is certainly not the *AGE* effect because the two samples were restricted to *AGE*=50. But is it the *PERIOD* effect or the *COHORT* effect? We don't know. Not only do we not know which true causal variables explain the decline, we can't even tell whether they are captured by *PERIOD* or by *COHORT*. For instance, did the prevalence decline because the cost of cigarettes was higher in 2010 than in 2000 (the *PERIOD* effect), or because anti-smoking messages were strengthened in the 1970s – the teenage years of the 1960 birth cohort (the *COHORT* effect)? Perhaps both. Perhaps neither. No, the oxymoron "hypothesis generating study" will not save the case. We don't need this study to propose and test theories about the effect of cigarette price on smoking, or the effect of anti-smoking messages.

The problem is not unique to my simple example. We cannot tell which effect is estimated even if *AGE*, *PERIOD*, and *COHORT* take many values. Suppose we try to estimate the three effects from the coefficients of a "main effects" regression model:

$$R = \beta_0 + \beta_1 A + \beta_2 P + \beta_3 C \qquad (1)$$

where *A, P,* and *C* denote *AGE*, *PERIOD*, and *COHORT*, respectively, and *R* is some response variable such as the probability of smoking.

AGE, *PERIOD*, and *COHORT* are not three independent variables, however. Any two of them fully determine the value of the third – a phenomenon called "perfect multicollinearity". For example,

$$A = P - C$$

Substituting *P-C* for *A* in (1), we get

$$R = \beta_0 + \beta_1 (P - C) + \beta_2 P + \beta_3 C$$

$$R = \beta_0 + (\beta_1 + \beta_2) P + (\beta_3 - \beta_1) C$$

Which means that only two parameters are actually specified (besides the intercept), not three:

$$R = \beta_0 + \gamma P + \delta C$$

But which two? We cannot identify them because the model may be written as a function of *any two* of the three variables.

From a mathematical standpoint there is no unique solution, because there are three unknown coefficients (β_1, β_2, β_3) and only two equations:

$$\beta_1 + \beta_2 = \gamma$$
$$\beta_3 - \beta_1 = \delta$$

Subtracting the second equation from the first, we derive a general formulation of what has been called "the identifiability problem of APC analysis":

$$2\beta_1 + \beta_2 - \beta_3 = \gamma - \delta$$

Any of the three coefficients is fully determined by the other two, or alternatively: the three coefficients may be written as a linear combination that sums up to a constant. The collinearity among the variables (C+A–P=0) resulted in collinearity among the coefficients too ($2\beta_1+\beta_2-\beta_3=\gamma-\delta$).

Do we need solutions?

The scientific problem of APC analysis cannot be solved because no argument can transform a theory that is thin in content into a rich content theory. At most, we may claim that APC analysis helps to predict future trends on the basis of past trends. Futurism of this type contains some elements of a primitive prediction model, but it may also be viewed as an extrapolation beyond the observed range of data (e.g., smoking prevalence in 2020), which is a questionable practice. For instance, past time trends may drastically change following the discovery of a new therapy, or the signing of a new law. Who can guess future discoveries or future laws?

Where the science is weak, statistical solutions are solutions of a math problem rather than aides to scientific discovery. The math problem – estimating coefficients under perfect multicollinearity – is relevant to science only if we declare interest in the coefficients of equation (1). Are they of interest indeed? What is their meaning?

In statistical jargon the boilerplate answer is well known: the coefficients tell us about "independent associations". For example, the coefficient of *PERIOD* quantifies the linear association of *PERIOD* with smoking, independent of *AGE* and *COHORT*. But anyone who understands the connection between expected associations and a causal structure knows that the magnitude of an independent association is not synonymous with the magnitude of an effect. In fact, there is no scientific merit in estimating an independent association, unless the breaking of dependencies serves to block unwanted open paths between the presumed cause-and-effect.[2]

Only if our theories are encoded in a causal diagram can we deduce which adjustment is needed to estimate an effect and which "adjustment" is redundant or possibly detrimental (increasing the bias, penalizing the variance, or both). Since the variables *AGE*, *PERIOD*, and *COHORT* do not show up in any causal structure, mutual adjustment is no more than a mathematical challenge, dressed up as a service to science.[a]

Moreover, under the axiom of indeterminism we don't expect perfect collinearity among natural variables (beyond chance collinearity in a given sample). Something is methodologically incoherent, if not ridiculous, in a problem that was created by choosing *AGE*, *PERIOD*, and *COHORT* to represent unknown causal variables among which the problem does not exist. Indeed, fit a model with true causal variables and the problem is gone.

Don't know which variables to model? Well, do you have a sharp theory to test, or just three variables that happen to be handy, along with pompous phrases such as "formative events in different eras" and "the lifetime experience of different birth cohorts"?

A load of solutions

I could have ended the commentary with the last section. Nothing more needs to be said. It is interesting, however, to see how statistical minds have tried to solve the identifiability problem of APC analysis. So let's pretend that *AGE*, *PERIOD*, and *COHORT* are causal variables; that causation is deterministic; and that mutual adjustment is indeed needed to estimate the effect of each variable.

Almost all of the solutions share the following idea: Replace equation (1) with another equation from which some parameters may be uniquely estimated – an estimable function, as it may be called. To that end, some constraint (restriction) must be imposed on the relation of the outcome with age, period, and cohort.[b]

Constraints of one kind reduce the number of parameters that need to be estimated (there is one too many). The remaining parameters are no longer perfectly collinear and may be uniquely estimated. Other constraints circumvent collinearity by *increasing* the number of parameters, as will be explained shortly. In scientific terms, all of the constraints are untestable theories about some effect(s).

Perhaps the simplest solution is categorization. The continuous variables are replaced with categorical variables, which are then modeled by dummy variables. As long as the length of the time interval is not identical for the categories of age, period and cohort, the new variables are no longer perfectly collinear; their coefficients can be estimated. What is the constraint? The effect is assumed to be identical for all values within each arbitrary-length interval (implying that only non-linear components are estimated[5]).

Interestingly, this solution of the identifiability problem goes against classic statistical reasoning. When a continuous exposure is categorized (say, to explore the dose-response function), we always strive for as narrow categories as possible – precisely because the effect is constrained to be identical within each category. The more data we have, the more we relax that constraint by choosing narrower categories (the limits of which are the values of the continuous variable). Here, no matter how much data are available, there is no motivation to shorten the

[a] Even under the weak theory of unspecified causal variables behind *COHORT* and *PERIOD*, there is no reason, for example, to adjust the association of smoking with *COHORT* for *PERIOD*. The association with a cohort variable (say, smoking in the high school years) is never confounded by a period variable (say, cigarette price in the survey year) because a confounder never follows the exposure on the time axis. Being a cause of the exposure, a confounder must *precede* the exposure. Unfortunately, most of the statisticians who wrote on APC analysis did not demonstrate clear understanding of the term "confounding bias" as formally derived from causal diagrams. Like many statisticians, they think that "independent association" (a statistical idea) and "unconfounded association" (a causal idea) are exchangeable terms. They are not.

[b] Formally, an estimable function is a function that provides unique estimates that *do not* depend on the chosen constraint. Here, I use the term liberally to denote any solution of the identifiability problem.

intervals and approach the continuous variable. Sounds awkward.

Another simple solution is a two-factor model. The number of independent variables is reduced from three to two by setting the coefficient of one variable to zero. For example, assume *a priori* that *PERIOD* has no effect on the outcome and thereby derive an estimable function:

$$R = \beta_0 + \beta_1 A + \beta_2 C \qquad (2)$$

On some occasions epidemiologists have been willing to adopt a two-factor model. For example, it is often claimed that the incidence of a disease is affected by *AGE* and by *COHORT*, but not by *PERIOD*. Of course, nothing is affected by any of these variables, but why can't some disease be prevented by a causal variable that is captured by *PERIOD* – say, a new drug? Moreover, does the method of science accept an untestable premise that one variable has a null effect on another?

Another possible constraint is equality of two coefficients (β_1=β_2 or β_2=β_3). For example, accept the untestable claim that the coefficient of *AGE* and the coefficient of *PERIOD* are identical:

$$R = \beta_0 + \beta_1 A + \beta_1 P + \beta_3 C \qquad (3)$$

Now we have a set of three equations, which has a unique solution:

$$\beta_1 - \beta_2 = 0$$
$$\beta_1 + \beta_2 = \gamma$$
$$\beta_3 - \beta_1 = \delta$$

(Notice that β_1=β_3, equality of *AGE* and *COHORT* effects, does not provide a unique solution.)

When all three variables are provided in equal length intervals, it is enough to assume equality of coefficients for two adjacent categories.[4-6] For instance, if the reference age group is 25-29, it is sufficient to assume that the coefficient of age group 30-34 is identical to the coefficient of age group 35-39. Again, perfect collinearity is avoided by reducing the number of estimated coefficients by one (or equivalently, by making one interval longer than all others).

Another type of constraint is to replace one of the three variables with a product of the other two. For example,

$$R = \beta_0 + \beta_1 A + \beta_2 P + \beta_3 AP \qquad (4)$$

which is a classic model of effect modification (interaction) between *AGE* and *PERIOD*. What happened to the cohort effect? Is it assumed to be null? The answer is tricky. If β_3=0 in expectation, the model is essentially reduced to a two-factor model, which implies a null effect of *COHORT*. Otherwise, the cohort effect is *redefined* as the phenomenon of effect modification between *AGE* and *PERIOD*. The identifiability problem remains, however. We cannot tell which two variables are the causes of the outcome and which variable should be redefined as the phenomenon of effect modification. Moreover, the model is irrelevant if effect modification operates only between coinciding causes.[7]

A different category of solutions calls for higher order polynomials, namely, for *more* parameters. For example,

$$R = \beta_0 + \beta_1 A + \beta_2 A^2 + \beta_3 P + \beta_4 P^2 + \beta_5 C + \beta_6 C^2 \quad (5a)$$

As before, the linear components of the function cannot be estimated due to collinearity of *A*, *P* and *C*, but the quadratic terms are not collinear. It is therefore possible to estimate the curvature of *R* as a function of each variable.[8] The merit of this solution is questionable, though: First, it is of no help if the functions happen to be strictly linear. Second, the curvatures do not add interesting knowledge, unless we also know the general (linear) direction of each effect.[9] Third, the model is not necessarily a three-factor model. It may also be written as a model of effect modification between two of the three variables. For example, if *C* is replaced with *P-A*, the model takes the following form:

$$R = \alpha_0 + \alpha_1 A + \alpha_2 A^2 + \alpha_3 P + \alpha_4 P^2 + \alpha_5 AP \quad (5b)$$

Which is just another way to specify effect modification between *A* and *P*. Again, which two variables are the causes of the outcome and which one should be redefined as the phenomenon of effect modification?

Partial least squares regression offers a unique solution of a linear regression model in cases of perfect multiple collinearity, such as equation (1). All three coefficients (β_1, β_2, and β_3) may be estimated by this method. How come? What is the magical solution for APC analysis? Buried in an appendix,[10] we find the answer: an implicit constraint on the coefficients (β_1+β_2=β_3). Some linear combinations of the coefficients underlie other functions that estimate esoteric parameters, such as differences between differences of unknown coefficients, and relationships between unknown slopes.[11]

Another peculiar solution is to replace the estimation of coefficients with the estimation of something else.[11] In ANOVA models (linear regression on dummy variables), the effect of age groups, period groups, and cohort groups is redefined as the contribution of the respective set of dummy variables to the variance of the outcome – after accounting for the other two sets. In other regression models, it is a likelihood ratio Chi-square – the improvement in model fit due to each set beyond the other two. What makes these statistics measures of effect, however? Does partitioning of the variance tell us about effects? Is any statistic that can be attached to a variable may be called "the effect of that variable"? Maybe so – in some minds.

When data are available for each person (or year), the identifiability problem may be solved by categorizing one of the three variables, followed by the fitting of a hierarchical mixed effects regression model.[c] For instance, create a categorical version of *COHORT* (by grouping adjacent birth cohorts); model the new variable as the cluster variable; and estimate the coefficients of *AGE* and *PERIOD*. What happened to the cohort effect, though? Well, its effect is now called "a random effect".

With mixed effects models at hand, the identifiability problem is easy to solve. Simply replace as many fixed effects coefficients as you wish – all of them, if possible – with random effects coefficients. Who needs those troublemakers anyway?

Here is one attempt to explain why we may categorize the *COHORT* variable (on the way to a hierarchical model):

"However, since meaningful cohorts often are considered to be of durations longer than single years, it then will be feasible to group the cohort dimension into multiyear periods while retaining single-year measurements for the age and time period dimensions."[12]

Translation: Since *COHORT* (birth year) is not really the variable of interest, its coefficient is not estimating any single causal parameter. The true causal variables behind the birth year are unspecified lifetime exposures, so exposures during any single year are not meaningful (unless they are captured by a single year of *PERIOD* or *AGE*?). As far as lifetime exposures are concerned, different cohort eras, which span several decades, are not that different if the non-overlapping time is relatively short – a few years. Therefore, we may safely combine adjacent birth cohorts (e.g., the birth years 1950-1955).

The hierarchical model approach, as outlined above, was thoroughly criticized both theoretically and empirically.[13-15] I will add only two comments: First, the argument for grouping adjacent birth cohorts holds for adjacent ages (*AGE*) and adjacent survey years (*PERIOD*), so any of the three variables may be designated as the cluster variable. Indeed, the preference for cohort (and period) in APC analysis was substantiated by "guidelines" – a dictatorial substitute for substantive arguments. Second, along this line of reasoning there is no justification for the retaining of *any* of the three continuous variables. All of them may be categorized simultaneously. In short, the hierarchical model approach looks like an attempt to force a problem on a statistical tool rather than finding a statistical tool that would fit a problem.

Finally, the so-called intrinsic estimator (IE) may be counted among the latest, popular solutions. Founded on principal component analysis,[d] the IE was claimed to be the ultimate solution to the identifiability conundrum. Too bad it turned out to be neither new nor unbiased nor constraint-free.[16,17]

A load of solutions indeed – almost too many to choose from. Fortunately, we don't need to worry about choosing any of them, because age, period, and cohort affect nothing and confound nothing. Their independent associations with any outcome are generic statistics, not scientific parameters. If you forgot why, read the previous section again.

Epilogue

Peculiarly enough, it was occasionally proposed to solve the identifiability problem by modeling a

[c] A mixed effects model contains terms for fixed effects (for example, the coefficients in equations 1-5) and terms for the so-called random effects. Mixed effects models are often fit when the data can be organized in clusters of correlated observations. Typical examples include repeated measurements, which are clustered – and correlated – in a person (longitudinal analysis); observations that are clustered in some group, such as students of a school or residents of a neighborhood (multi-level analysis); and observations that are clustered in a study (meta-analysis). These models are often called "hierarchical" because the cluster variable is "above" the individual observations (and there may be a cluster of clusters). In mixed effects APC models, any of the three variables may be considered the cluster variable (for which no fixed-effect coefficient is estimated). Since there is no natural hierarchy here, the adjective "hierarchical" is not a correct descriptor of these models.

[d] Principal component analysis transforms a set of highly correlated variables into a set of new, linearly uncorrelated variables, which are called principal components. The latter are created hierarchically according to the remaining variance (largest to smallest) that is accounted for by each new variable. The number of principal components may be smaller than the number of the original variables.

subject matter variable, along with *COHORT*, that might “explain” some of the cohort effect. Perhaps more than anything, this solution illustrates the “upside down” thinking about the scientific matter. Such a variable does not explain any “cohort effect”. It is the only variable in the model that might have an effect on the outcome!

The so-called age effect, period effect, and cohort effect are not effects at all in causal reality. Therefore, the problem of estimating independent associations in APC analysis is no more than a mathematical challenge, mistakenly portrayed as relevant to science. Analysts need not worry about choosing a model for APC analysis.[11,18] Whichever independent association they are trying to estimate, it is not an estimate of a causal parameter. As for quantitative time trends, they have more in common with good history (the study of the past) than with good science (the study of causal laws).

Acknowledgement: Doron Shahar – for comments on a draft manuscript

References

1. Shahar E, Shahar DJ. Causal diagrams, information bias, and thought bias. *Pragmatic and Observational Research* 2010;1:33–47
2. Shahar E, Shahar DJ: Causal diagrams and three pairs of biases. In: *Epidemiology – Current Perspectives on Research and Practice* (Lunet N, Editor). www.intechopen.com/books/epidemiology-current-perspectives-on-research-and-practice, 2012:pp. 31-62
3. Popper KR. *The Logic of Scientific Discovery*. Hutchinson Education, 1959; Routlage, 1992
4. Mason, KO, Mason WM, Winsborough HH, Poole WK. Some methodological issues in cohort analysis of archival data. *American Sociological Review* 1973;38:242-58
5. Fienberg SE, Mason WM, Identification and estimation of age-period-cohort models in the analysis of discrete archival data. *Sociological Methodology* 1979;10:1-67
6. Kupper LL, Janis JM, Karmous A, Greenberg BG. Statistical age-period-cohort analysis: a review and critique. *Journal of Chronic Diseases* 1985;38:811-830
7. Shahar E. On effect modification and its applications. http://www.u.arizona.edu/~shahar/commentaries.html
8. Holford TR. An alternative approach to statistical age-period-cohort analysis. *Journal of Chronic Diseases* 1985;38:831-836
9. Authos’ reply. *Journal of Chronic Diseases* 1985;38:837-840
10. Tu Y-K, Davey Smith G, Gilthorpe MS. A new approach to age-period-cohort analysis using partial least squares regression: the trend in blood pressure in the Glasgow Alumni cohort. *PLoS One* 2011 Apr 27;6(4):e19401. doi: 10.1371/journal.pone.0019401
11. O'Brien R. *Age-Period-Cohort Models: Approaches and Analyses with Aggregate Data* [e-book]. Boca Raton, FL: CRC Press; 2014. Available from: eBook Collection (EBSCOhost), Ipswich, MA
12. Yang Y, Land KC. Age-period-cohort analysis of repeated cross-section surveys: fixed or random effects? *Sociological Methods & Research* 2008;36:297-326
13. Bell A, Jones K. Don't birth cohorts matter? A commentary and simulation exercise on Reither, Hauser and Yang's (2009) age-period-cohort study of obesity. *Social Science & Medicine* 2014; 101:176-180
14. Bell A, Jones K. Another 'futile quest'? A simulation study of Yang and Land's hierarchical age-period-cohort model. *Demographic Research* 2014;30:333-360
15. Bell A, Jones K. Should age-period-cohort analysts accept innovation without scrutiny? A response to Reither, Masters, Yang, Powers, Zheng and Land. *Social Science & Medicine* 2015;128:331-333
16. Luo L. Assessing validity and application scope of the intrinsic estimator approach to the age-period-cohort problem. *Demography* 2013;50:1945-1967
17. Luo L. Paradigm shift in age-period-cohort analysis: a response to Young and Land, O’Brien, Held and Riebler, and Fienberg. *Demography* 2013;50:1985-1988
18. Yang Y, Land KC. *Age-Period-Cohort Analysis: New Models, Methods, and Empirical Applications*. Chapman & Hall/CRC, 2013