Estimation with Vanishing Baseline Risk

Robert M. Park

Centers for Disease Control and Prevention, National Institute for Occupational Safety and Health Education and Information Division, Columbia Parkway, Cincinnati, OH, rhp9@cdc.gov

To the Editor

Diseases associated with specific exposures may have little or no observable background rate in the absence of the exposure. Examples include mesothelioma (environmental asbestos), aplastic anemia (benzene), bronchiolitis obliterans (artificial butter flavorings), Reye's syndrome (aspirin in children), and angiosarcoma of the liver (vinyl chloride). Relative-rate models of exposure-response produce unstable near-zero baseline risk and unbounded coefficients, especially when age confounding requires baseline age dependence. The same problem arises in a proportional-hazards context. Baseline risk volatility also threatens meta-analyses, a procedure that assumes uniformity.

Using Poisson regression,[1] we investigated two methods: (1) fixing the intercept at a small value corresponding to 1% of attributable cases and (2) generating random sets of new cases across observation time independent of any predictor, possibly preempting true cases. Although models can be reliably fit using randomly generated cases, repetition would reduce variability in parameter estimates. We performed simulations with fixed intercepts (1,000) and with simulated populations (100) each with 100 random baselines. Hypothetical populations, constructed iteratively, consisted of 500 subjects with an exposure that could extend up to 200 time units. Exposure duration was random, favoring shorter durations to represent typical environmental or occupational exposures. Individual average exposure levels were randomly assigned and then randomly varied across time. We generated attributable cases with probability proportional to cumulative exposure, at which time follow-up ceased. Numbers of attributable or baseline cases averaged ~60–70. The analyses were implemented using an R algorithm[2] that called specific FORTRAN and EPICURE[3] steps with an indexing seed for random number generation. Additional information is included in the eAppendix (http://links.lww.com/EDE/A619).

The model specification was as follows:

$$\text{rate} = [\exp(\alpha)] \times [1 + \beta \text{cumX}] \text{ or rate ratio} = 1 + \beta \text{cumX},$$

where cumX is an exposure metric, α is the intercept defining baseline risk, βcumX is the excess rate ratio, and β is the excess rate ratio coefficient.

Analyses were conducted as follows:

1. Attributable cases only

2. Attributable cases only, analyzed with fixed intercept

3. With added nonattributable cases

4. With added nonattributable cases analyzed with intercept fixed at known baseline risk (number of baseline cases/person-years of observation).

With the standard model, the excess rate ratio coefficient, β, varied widely across 1000 populations: mean = 13.4 (SD = 94.5) and range = 0.1-2834; with constrained intercept, the mean = 5.9 (0.76); range = 3.7-8.8. The mean of log(excess rate ratio coefficient) was 1.54 (SD =1.3) versus 1.77 (0.13) with fixed intercept (Table). The mean excess rate coefficient, $\exp(\alpha)] \times \beta$, nominally 0.00006 in the simulation, was close to nominal with fixed intercepts (0.00005981), but biased downward in standard models (0.00005095) by 15%. The mean-squared deviation of the excess rate coefficient was substantially smaller with fixed intercepts (0.59×10^{-10}) versus standard model (1.62×10^{-10}), a 63% reduction.

In 100 simulated populations, each with 100 iterations of added baseline cases, estimates of excess rate ratio coefficient were much less variable than with standard models, especially with intercept fixed at the known baseline risk. The mean excess rate coefficient was now close to nominal with or without the fixed intercept (0.00005964 and 0.00006009, respectively). When the average squared deviation of the estimated excess rate coefficient was calculated within each set of 100 baseline iterations, the mean of those averages across the 100 simulated populations with intercepts fixed (0.45×10^{-10}), was comparable to that without baseline enhancements but with fixed intercepts (0.59×10^{-10}).

Simulations with small populations (n = 50) demonstrated greater bias (Table). The excess rate coefficient bias was 15% and 32% in the populations with 500 and 50 subjects, respectively. The two treatments for vanishing baseline yield equivalent results demonstrating that simply fixing the intercept is entirely adequate.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

REFERENCES

1. Frome EL, Checkoway H. Epidemiologic programs for computers and calculators. Use of Poisson regression models in estimating incidence rates and ratios. Am J Epidemiol. 1985; 121:309–323. [PubMed: 3839345]

2. Venables, WN.; Smith, DM. An Introduction to R. Bristol, UK: Network Theory; 2002. Available at: http://www.r-project.org/. [Accessed April 4, 2010]

3. Hirosoft International Corporation. Epicure Users Guide. Seattle, WA: Hirosoft International Corporation; 1993.

TABLE

Summary Comparisons of Estimation Performance With and Without Fixed Intercept or Random Baseline for Large and Small Population Simulations

	Large Sample (500)		Small Sample (50)	
	Mean	**(SD)**	**Mean**	**(SD)**
Log(excess rate ratio coefficient), log(β)				
Estimated intercept/no baseline (n = 1,000)	1.54	(1.34)	0.63	(1.86)
Fixed intercept/no baseline (n = 1,000)	1.77	(0.13)	1.69	(0.44)
Excess rate coefficient, [exp(α)] $\times \beta$ ($\times 10^5$), nominal value = 6.000				
Estimated intercept/no baseline (n = 1,000)	5.095	(0.90)	4.082	(2.19)
Fixed intercept/no baseline (n = 1,000)	5.981	(0.77)	5.989	(2.40)
Fixed intercept/random baseline, avg (n = 100[a])	5.964	(0.67)	6.278	(2.71)
Squared deviation: (excess rate coefficient − 6.0×10^{-5})^2 ($\times 10^{10}$)				
Estimated intercept/no baseline (n = 1,000)	1.62	(1.77)	8.48	(9.63)
Fixed intercept/no baseline (n = 1,000)	0.59	(0.87)	5.76	(9.33)
Fixed intercept/random baseline, avg (n = 100)	0.45	(0.46)	7.34	(14.5)

SD indicates standard deviation.

[a]Based on 100 iterations of study population each analyzed with 100 random baselines; average for each study population across the set of its 100 random baselines.