May 31, 2025

Below you will find our (twice-reviewed but unpublished) work on **sample size considerations for case crossover studies**. It was prompted by questions my colleague Scott Weichenthal had when he was planning case-crossover studies in the field of environmental epidemiology.

But, of course, these considerations also apply to case-crossover studies in other fields. I was also keen to play up the connections I saw with the Cox model, which shares the same likelihood function with the model used in conditional (matched set) logistic regression.

We submitted it first as a tutorial (in 'epidemiologic statistics') in a broader and more methodologic journal (IJE, the International Journal of Epidemiology), and then to an environmental research journal. Both submissions included an extensive appendix that showed the broader connections, and the insights from inspecting the likelihood function.

Below you can read

- an early draft, which included at the end some additional notes, possible references, and examples.
- The submitted versions, and the reviews we received.

We still think there are important (and broader) messages/insights in this manuscript.

In section 10 of the supplement to my recent (published) article[1] on a very early application of conditional logistic regression – 4 decades before the dates one sees when one looks up this topic in Wikipedia – I commented on the form that the variance of the fitted coefficients takes. Note how critical it is to have a large within-set variance of the exposure of interest. Also, if the (again within-set) covariance with any 'confounding' variable is high, one would need a large number of sets to counteract the smaller 'effective variance' for the exposure of interest – and the resulting imprecision. Penrose's mentor, R A Fisher, remarked on this in their extensive correspondence.

Even though I don't have any current plans to see this manuscript through to publication, I would be happy to help a younger and more energetic person bring its messages to a broader readership. So, if you are interested, please contact me.

Sincerely,
James Hanley

webpage: https://jhanley.biostat.mcgill.ca

email: `james.hanley@mcgill.ca`

[1] https://jhanley.biostat.mcgill.ca/Penrose/

Draft 2021.05.27

The planning, analysis and interpretation of 'case crossover' studies involving a quantitative environmental 'exposure'

Scott Weichenthal[1 2],   xxx?   James A. Hanley[1]

[1]Department of Epidemiology, Biostatistics, and Occupational Health,
[2] Department of Oncology

McGill University, Montreal, Canada

Keywords:
parameter estimation; Poisson regression; conditional logistic regression; connections, sample size; exposure variance

Abstract

Background: Reports of ‘case crossover’ studies linking the rate of health events to a quantitative environmental ‘exposure’ are becoming much more common, but the statistical aspects behind the planning and analysis are not very transparent or intuitive.

Methods: Using a small hand-worked example, we illustrate the Maximum Likelihood (ML) calculations involved in estimating the parameter of the conditional logistic regression model that quantifies the strength of the exposure-response function. The quantity that determines the precision of the parameter estimate is easily seen from these calculations and provides a direct way to anticipate the precision and statistical power that will result from a given sample size. The amount of statistical ‘information’ each instance contributes to the ML parameter estimate is emphasized.

Results/Conclusions: As expected, the standard error of the estimated regression coefficient estimate is inversely related to the square root of the number of instances (cases) of the event. It is also inversely related to a function of the variation in the exposure values in a typical matched set. This understanding, and the fact that the statistical power depends on two versions of this typical variation (one under the ‘null’ and one under the ‘alternative’) can be used to plan the size of a case-crossover study. A study involving of an all-or-none exposure is just a special case.

218 words

Introduction

Articles with abstracts stating that the authors 'conducted a case-crossover study of xxx health events using weather records between the months of xxx and yyy for place z from 19xx to 20yy' are becoming more and more common. Authors typically report that they 'compared some measure of exposure on/before the date of these events with the same measure on (possibly matched) control days when the event did not occur.' They also typically report that they 'used conditional logistic regression to compute odds ratios (OR) and 95% confidence intervals (CI) to measure the association between the exposure and event of interest.'

Unfortunately, guidance on the statistical aspects behind the planning and analysis is not as transparent or intuitive as it might be Thus, we use a small hand-worked example to show the logic behind, and the calculations involved in obtaining, the Maximum Likelihood (ML) estimate the parameter of the conditional logistic regression model that quantifies the strength of the exposure-response function. We also illustrate the use of Poisson regression models. We examine the 'anatomy' of the formula for the precision of the parameter estimate. The resulting insights provide a direct way to anticipate the precision and statistical power that will result from a given sample size. Throughout, we emphasize the amount of statistical 'information' each instance contributed to the ML parameter estimate.

Even though our focus is on quantitative exposures, the exact same statistical principles apply to a case-crossover study involving an all-or-none exposure. Since it is merely a special case, the two statistical silos can be unified under the single unified approach presented here, just as was done earlier (Hanley and Moodie, 2011). Indeed, if viewed appropriately, the formula given under the Rate Ratios section of that article applies to case-cross-over studies as well.

In our review of recent so-called case-crossover studies, have noticed some confusion in terminology, 'creep' from the original case-crossover studies, and misunderstandings as what parameter of the exposure-response relationship is being estimated by the conditional logistic regression. We also see multiplicity issues involving lagged exposures. Thus, we address these in the Discussion

A small, but real, dataset for illustration

Figure 1 shows a small, but real, dataset typical of many environmental case-crossover studies. The heat map shows the daily maximum temperatures. Since they are from a North American setting, they range from very cold (dark blue) in Winter to very hot (dark red) in Summer. The dates of the 10 events of interest are indicated by rectangles with solid black borders: they are confined to the months from May to September.

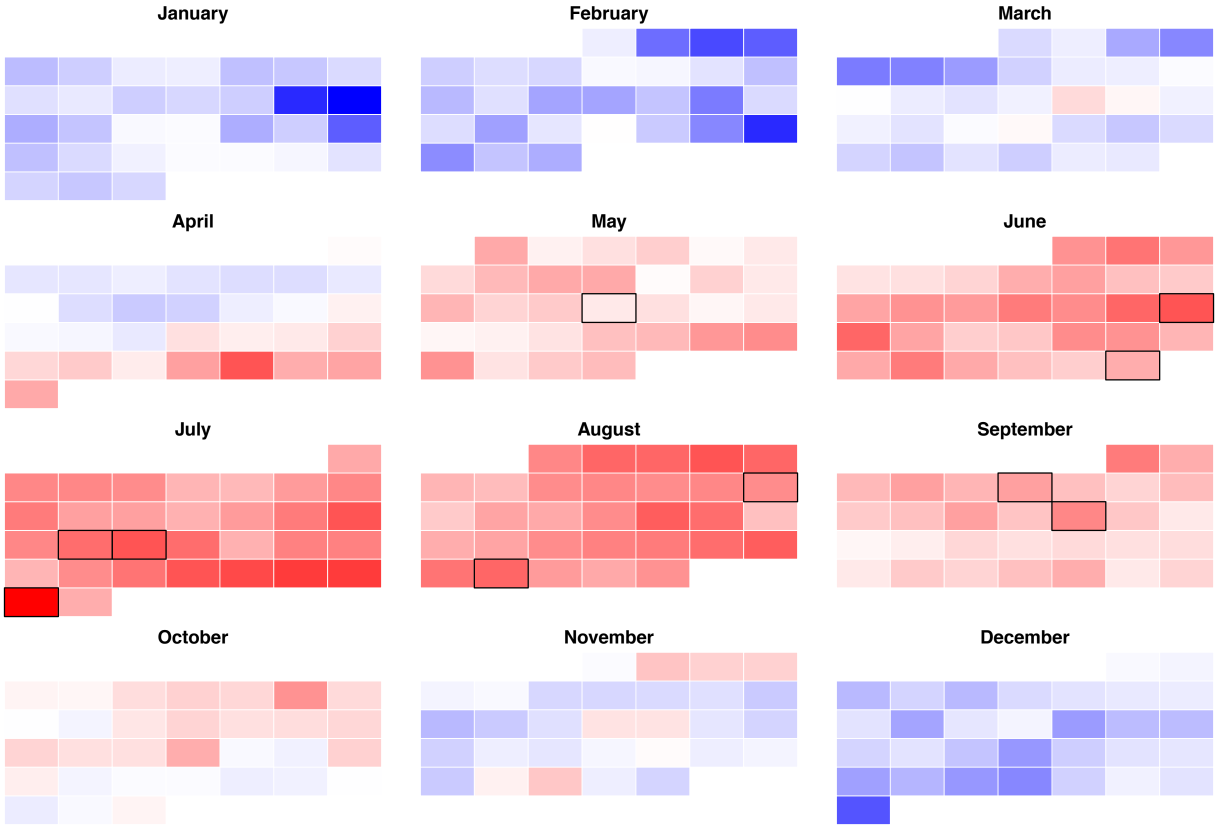


Figure 1*: A small, but real, dataset bearing on the relationship between the rate of an event of concern and the daily temperature. The dates of the 10 events are marked by solid black borders, and the magnitudes of the daily temperatures are color-coded from cold (blue) to hot (red).*

Before moving on to the data analysis, we comment briefly in passing on terminology, a topic we will return to later. Imagine for a moment that this Figure did not yet contain the dates of the events, but did contain the temperatures. Then, with the planned matching on the month and day of the week, the design can be thought of as a total of 7 x 12 = 84 *possible* mini-comparisons or regression models, one for each of the 84 'columns' of 4 or 5 days. As we will see in the next sections, even if the plan is to include all 365 data points in a regression model in which the rate is taken to be log-linear in temperature within each matched column, then only the

44 data points from the 10 columns containing at least one event will contribute to the parameter estimate of interest. Nevertheless, this 'before-seeing-the-event-data' regression plan/outlook still makes it clear that – conceptually at least -- we *are comparing event rates across a range of temperatures;* we are *not comparing* '*cases' with* '*controls'* (or 'case days' versus 'control' or 'reference' days) with respect to temperature (exposure). Whereas the term 'case-crossover' may be familiar to epidemiology researchers, it does not communicate to a lay consumer that what are really being compared (with some additional time-matching) are *event rates at various temperatures*.

The exposure-response model

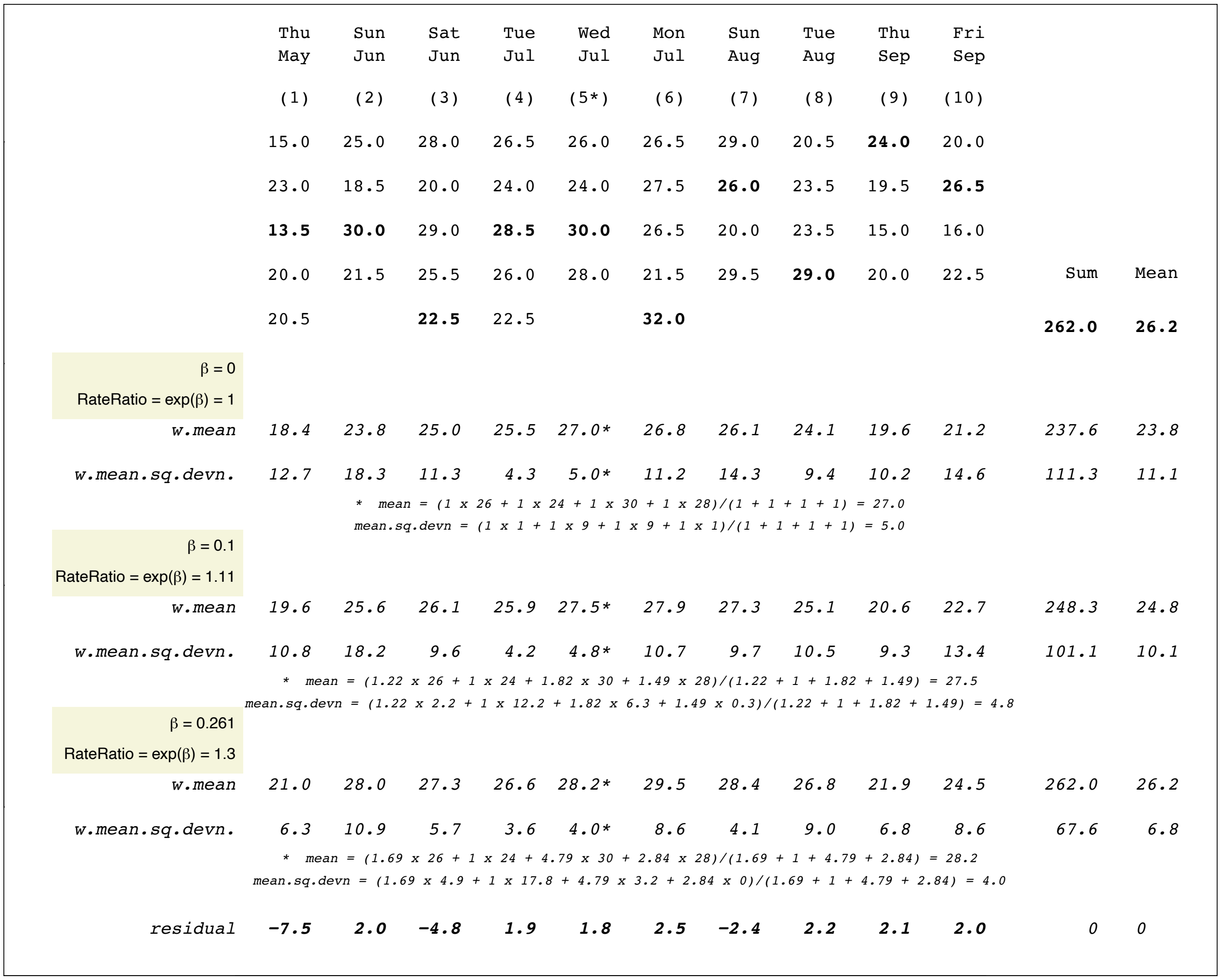
Without loss of generality, we will limit ourselves to the temperature on the day (T), rather than to any lagged version of it, and to the usual multiplicative model for event rates i.e., for the expected numbers of events per day

$$\lambda(T) = \lambda_0 \times \exp[\,\beta\,(\mathrm{T} - T_0\;)\,]\,,$$

where λ_0 refers to the event rate at some reference temperature, T_0, the shorthand 'exp' stands for 'the exponential of' and, thus, $\exp[\,\beta\,(\mathrm{T} - T_0)\,]$ denotes the ratio of the event rate at temperature T to the rate at the reference temperature (to simplify matters, we will take T_0 to be 0° Celsius, but will come back to this when addressing the reporting of model fits). Thus, β, the parameter of interest, refers to the log of the ratio of the event rates at temperatures that are 1° Celsius apart. Despite the fact that this rate ratio model is usually *fitted* via a *logistic* regression, its exponentiated value should be referred to as a *rate* ratio rather than an *odds* ratio.

The fitting of this model to the illustrative dataset

The top 4/5 rows of Figure 2 show the temperatures for the day-of-week-that-month 'column' for each of the 10 events.

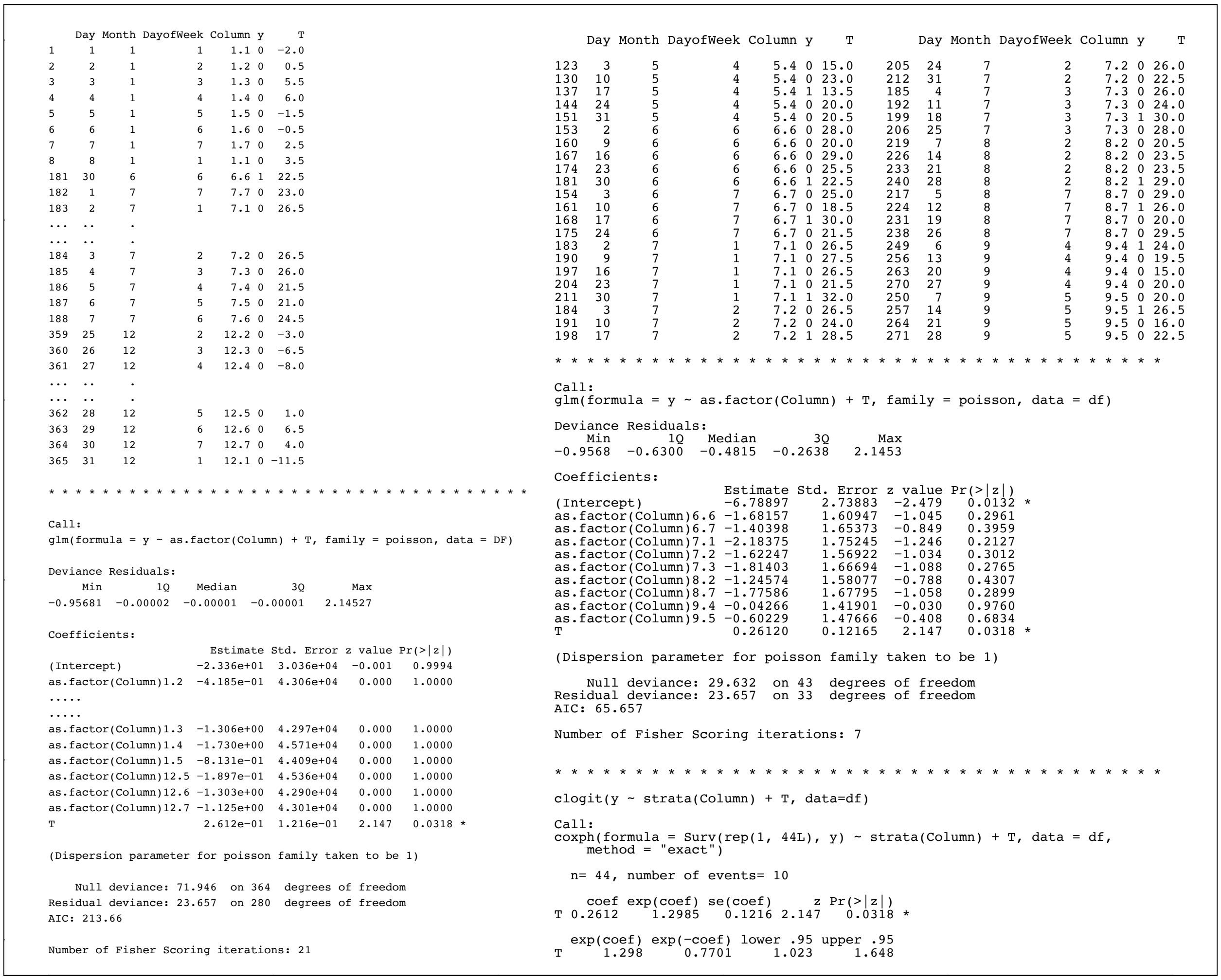


	Thu May	Sun Jun	Sat Jun	Tue Jul	Wed Jul	Mon Jul	Sun Aug	Tue Aug	Thu Sep	Fri Sep		
	(1)	(2)	(3)	(4)	(5*)	(6)	(7)	(8)	(9)	(10)		
	15.0	25.0	28.0	26.5	26.0	26.5	29.0	20.5	**24.0**	20.0		
	23.0	18.5	20.0	24.0	24.0	27.5	**26.0**	23.5	19.5	**26.5**		
	13.5	**30.0**	29.0	**28.5**	**30.0**	26.5	20.0	23.5	15.0	16.0		
	20.0	21.5	25.5	26.0	28.0	21.5	29.5	**29.0**	20.0	22.5	Sum	Mean
	20.5		**22.5**	22.5		**32.0**					**262.0**	**26.2**
β = 0 RateRatio = exp(β) = 1												
w.mean	*18.4*	*23.8*	*25.0*	*25.5*	*27.0**	*26.8*	*26.1*	*24.1*	*19.6*	*21.2*	*237.6*	*23.8*
w.mean.sq.devn.	*12.7*	*18.3*	*11.3*	*4.3*	*5.0**	*11.2*	*14.3*	*9.4*	*10.2*	*14.6*	*111.3*	*11.1*
** mean = (1 x 26 + 1 x 24 + 1 x 30 + 1 x 28)/(1 + 1 + 1 + 1) = 27.0* *mean.sq.devn = (1 x 1 + 1 x 9 + 1 x 9 + 1 x 1)/(1 + 1 + 1 + 1) = 5.0*												
β = 0.1 RateRatio = exp(β) = 1.11												
w.mean	*19.6*	*25.6*	*26.1*	*25.9*	*27.5**	*27.9*	*27.3*	*25.1*	*20.6*	*22.7*	*248.3*	*24.8*
w.mean.sq.devn.	*10.8*	*18.2*	*9.6*	*4.2*	*4.8**	*10.7*	*9.7*	*10.5*	*9.3*	*13.4*	*101.1*	*10.1*
** mean = (1.22 x 26 + 1 x 24 + 1.82 x 30 + 1.49 x 28)/(1.22 + 1 + 1.82 + 1.49) = 27.5* *mean.sq.devn = (1.22 x 2.2 + 1 x 12.2 + 1.82 x 6.3 + 1.49 x 0.3)/(1.22 + 1 + 1.82 + 1.49) = 4.8*												
β = 0.261 RateRatio = exp(β) = 1.3												
w.mean	*21.0*	*28.0*	*27.3*	*26.6*	*28.2**	*29.5*	*28.4*	*26.8*	*21.9*	*24.5*	*262.0*	*26.2*
w.mean.sq.devn.	*6.3*	*10.9*	*5.7*	*3.6*	*4.0**	*8.6*	*4.1*	*9.0*	*6.8*	*8.6*	*67.6*	*6.8*
** mean = (1.69 x 26 + 1 x 24 + 4.79 x 30 + 2.84 x 28)/(1.69 + 1 + 4.79 + 2.84) = 28.2* *mean.sq.devn = (1.69 x 4.9 + 1 x 17.8 + 4.79 x 3.2 + 2.84 x 0)/(1.69 + 1 + 4.79 + 2.84) = 4.0*												
residual	***−7.5***	***2.0***	***−4.8***	***1.9***	***1.8***	***2.5***	***−2.4***	***2.2***	***2.1***	***2.0***	*0*	*0*

***Figure 2**: Top: the temperatures, T, (° Celsius) for the day-of-week-that-month 'column' containing each of the 10 events shown in Figure 1. The temperature on the day of the event is indicated in bold.*
Bottom: the calculations used in the pursuit of the Maximum Likelihood (ML) estimate of β, starting with the null value. Each `mean` *is a weighted average of the temperatures $T_1, T_2, \ldots\ T_{4/5}$, with weights $\exp[\beta T_1]$, $\exp[\beta T_2]$, ... or, equivalently, as shown, with re-scaled weights $\exp[\beta T'_1]$, $\exp[\beta T'_2]$, ... where $T'_1, T'_2, \ldots\ T'_{4/5}$ are measured relative to the minimum T in the column, Thus, the minimum temperature in the column has a weight of 1. Each* `mean.sq.devn` *is a weighted average of the squared deviations of $T_1, T_2, \ldots\ T_{4/5}$ from* `mean`*, using these same weights. The detailed calculations are shown for one selected column (5*). The sum/mean at the right is the sum/mean over the 10 instances/cases. The ML iterations continue until the sum/mean of the 10 fitted/weighted means equals (balances) the sum/mean of the 10 (observed) temperatures on the days the events occurred.*

Later we will use the calculations in the bottom portion of Figure 2 to show how and why conditional logistic regression can be used to fit the parameter of interest, β, in the exposure-response model. But first, to appreciate why a *conditional* approach is preferred, we begin by showing how it can be fitted using a much more familiar event-rate model, namely a Poisson regression model in which the time-matching is dealt with in the model, via indicator ('dummy') variables, one per column. Since each of the rows involves 1 day's duration, the offset (the log of

1 day, or 0) can be omitted. Since the total number of events is just 10, we would expect that the coefficients for the indicator variables will be very imprecisely estimated. And indeed, as is shown in Figure 3, they are. But more remarkable, and *less well known*, is that the β coefficient for T (temperature) has a standard error that is commensurate with the number of events, and is not distorted (biased) by the overfitting of such a large model to so little data (just 10 events). [The model on the right had side of Figure 3 has 84 + 1 = 85 terms, while the one of the upper left has 10 + 1 = 11]. This is a little-known feature that is peculiar to the *Poisson* model (it is not true in the case of a logistic regression that has as many intercepts as there are matched sets --see Breslow and Day, Volume I, page xx). The fitted β of 0.261 implies that the event rate is exp[0.261] = 1.3 times higher at (x + 1) °C than it is at x °C.



```
     Day Month DayofWeek Column y      T
1      1     1         1    1.1 0   -2.0
2      2     1         2    1.2 0    0.5
3      3     1         3    1.3 0    5.5
4      4     1         4    1.4 0    6.0
5      5     1         5    1.5 0   -1.5
6      6     1         6    1.6 0   -0.5
7      7     1         7    1.7 0    2.5
8      8     1         1    1.1 0    3.5
181  30     6         6    6.6 1   22.5
182   1     7         7    7.7 0   23.0
183   2     7         1    7.1 0   26.5
...  ..     .
...  ..     .
184   3     7         2    7.2 0   26.5
185   4     7         3    7.3 0   26.0
186   5     7         4    7.4 0   21.5
187   6     7         5    7.5 0   21.0
188   7     7         6    7.6 0   24.5
359  25    12         2   12.2 0   -3.0
360  26    12         3   12.3 0   -6.5
361  27    12         4   12.4 0   -8.0
...  ..     .
...  ..     .
362  28    12         5   12.5 0    1.0
363  29    12         6   12.6 0    6.5
364  30    12         7   12.7 0    4.0
365  31    12         1   12.1 0  -11.5

* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *

Call:
glm(formula = y ~ as.factor(Column) + T, family = poisson, data = DF)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.95681  -0.00002  -0.00001  -0.00001   2.14527

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)           -2.336e+01  3.036e+04  -0.001   0.9994
as.factor(Column)1.2  -4.185e-01  4.306e+04   0.000   1.0000
.....
.....
as.factor(Column)1.3  -1.306e+00  4.297e+04   0.000   1.0000
as.factor(Column)1.4  -1.730e+00  4.571e+04   0.000   1.0000
as.factor(Column)1.5  -8.131e-01  4.409e+04   0.000   1.0000
as.factor(Column)12.5 -1.897e-01  4.536e+04   0.000   1.0000
as.factor(Column)12.6 -1.303e+00  4.290e+04   0.000   1.0000
as.factor(Column)12.7 -1.125e+00  4.301e+04   0.000   1.0000
T                      2.612e-01  1.216e-01   2.147   0.0318 *

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 71.946  on 364  degrees of freedom
Residual deviance: 23.657  on 280  degrees of freedom
AIC: 213.66

Number of Fisher Scoring iterations: 21
```

```
    Day Month DayofWeek Column y    T          Day Month DayofWeek Column y    T

123   3     5         4    5.4 0 15.0     205  24     7         2    7.2 0 26.0
130  10     5         4    5.4 0 23.0     212  31     7         2    7.2 0 22.5
137  17     5         4    5.4 1 13.5     185   4     7         3    7.3 0 26.0
144  24     5         4    5.4 0 20.0     192  11     7         3    7.3 0 24.0
151  31     5         4    5.4 0 20.5     199  18     7         3    7.3 1 30.0
153   2     6         6    6.6 0 28.0     206  25     7         3    7.3 0 28.0
160   9     6         6    6.6 0 20.0     219   7     8         2    8.2 0 20.5
167  16     6         6    6.6 0 29.0     226  14     8         2    8.2 0 23.5
174  23     6         6    6.6 0 25.5     233  21     8         2    8.2 0 23.5
181  30     6         6    6.6 1 22.5     240  28     8         2    8.2 1 29.0
154   3     6         7    6.7 0 25.0     217   5     8         7    8.7 0 29.0
161  10     6         7    6.7 0 18.5     224  12     8         7    8.7 1 26.0
168  17     6         7    6.7 1 30.0     231  19     8         7    8.7 0 20.0
175  24     6         7    6.7 0 21.5     238  26     8         7    8.7 0 29.5
183   2     7         1    7.1 0 26.5     249   6     9         4    9.4 1 24.0
190   9     7         1    7.1 0 27.5     256  13     9         4    9.4 0 19.5
197  16     7         1    7.1 0 26.5     263  20     9         4    9.4 0 15.0
204  23     7         1    7.1 0 21.5     270  27     9         4    9.4 0 20.0
211  30     7         1    7.1 1 32.0     250   7     9         5    9.5 0 20.0
184   3     7         2    7.2 0 26.5     257  14     9         5    9.5 1 26.5
191  10     7         2    7.2 0 24.0     264  21     9         5    9.5 0 16.0
198  17     7         2    7.2 1 28.5     271  28     9         5    9.5 0 22.5

* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *

Call:
glm(formula = y ~ as.factor(Column) + T, family = poisson, data = df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9568  -0.6300  -0.4815  -0.2638   2.1453

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)          -6.78897    2.73883  -2.479   0.0132 *
as.factor(Column)6.6 -1.68157    1.60947  -1.045   0.2961
as.factor(Column)6.7 -1.40398    1.65373  -0.849   0.3959
as.factor(Column)7.1 -2.18375    1.75245  -1.246   0.2127
as.factor(Column)7.2 -1.62247    1.56922  -1.034   0.3012
as.factor(Column)7.3 -1.81403    1.66694  -1.088   0.2765
as.factor(Column)8.2 -1.24574    1.58077  -0.788   0.4307
as.factor(Column)8.7 -1.77586    1.67795  -1.058   0.2899
as.factor(Column)9.4 -0.04266    1.41901  -0.030   0.9760
as.factor(Column)9.5 -0.60229    1.47666  -0.408   0.6834
T                     0.26120    0.12165   2.147   0.0318 *

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 29.632  on 43  degrees of freedom
Residual deviance: 23.657  on 33  degrees of freedom
AIC: 65.657

Number of Fisher Scoring iterations: 7


* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *

clogit(y ~ strata(Column) + T, data=df)

Call:
coxph(formula = Surv(rep(1, 44L), y) ~ strata(Column) + T, data = df,
    method = "exact")

  n= 44, number of events= 10

    coef exp(coef) se(coef)     z Pr(>|z|)
T 0.2612    1.2985   0.1216 2.147   0.0318 *

  exp(coef) exp(-coef) lower .95 upper .95
T     1.298     0.7701     1.023     1.648
```

***Figure 3**: <u>left</u>: The data for each of the 365 days (for lack of space, only 24 are shown), followed by the Poisson regression model, with 12 x 7 = 84 indicator variables, one per Month-DayOfWeek. y=1 denotes an event. The row*

for the coefficient of T, 2.61 10^{-1} or 0.2612 shows the fitted β. The standard errors for the coefficients of the 84 indicator variables are all very large, because the total number of events is just 10.
Right: The data for each of the 44 days for the 10 Month-DayOfWeek combinations in which an event occurred, followed by the fitted Poisson regression model, with 10 indicator variables, one per Month-DayOfWeek. The bottom right portion shows the fitted conditional logistic model. All three regression approaches yield an identical coefficient of T, 0.2612, for the fitted β, as well as the identical standard error of 0.1216.

Clearly, it would be preferable if we did not have to rely on this peculiarity of the Poisson model, and if we could instead use a smaller model in which the matching in the analysis was accomplished by real rather than by rmodel-based matching. As Chapters 13 and 15 of Clayton and Hills elegantly show, the matching in the analysis can sometimes be accomplished by treating the total number of events within the matched set as a *fixed* quantity rather than the random variable that it is, and thereby eliminating the 84/10 intercepts, which represent nuisance parameters. (This is the same approach to nuisance parameters that is used in Fisher's exact test). In the examples addressed in these chapters, how the events distribute themselves within the 'exposed' and 'unexposed' person-time can be described by a *bi*nomial random carriable, in which the number of events serves as the '*n*' and the probability parameter is a function of the amounts of person time and the rate ratio. In our context, where the experimental units are days within a column, and exposure each day is measured on a quantitative scale, how the events distribute themselves over the possible days can be described by a *multi*nomial random variable, in which the number of events serves as the '*n*' and the multinomial probabilty parameters are a function of the amounts of person time and the rate ratios. For example, in column (5*) in Figure 2, the temperatures on the 4 candidate days are 26, 24, 30 and 28 °C. Thus, *given that* an event occurred on one of these days (i.e., *conditional* on the event having occurred within the column), the multinomial probabilities that it occurred on the first, second third and fourth of these days are, respectively,

$$\frac{\{\,exp[26\beta],\;\; exp[24\beta],\;\; exp[30\beta],\;\; exp[28\beta]\,\}}{exp[26\beta] + exp[24\beta] + exp[30\beta] + exp[28\beta]}\,.$$

In 2000, the Nobel Prize in economics was awarded to Daniel McFadden for his refinement of this model for his microeconometric analyses of choice behavior of consumers who face discrete ecomonic alternatives. In his Nobel lecture, McFadden tells how in 1965 he "called this a conditional *logit* model, since in the case of binomial choice it reduced to the logistic model used in biostatistics" but he used the *multinomial logit* (MNL) model terminology that is more common in economics today. In 2002, Norman Breslow devoted much of his address to the International Biometric Conference to the parallel developments of this model in epidemiology and biostatistics. In a very instructive article, Pardoe and Simonton used this discrete choice model to predict Academy Award winners.

McFadden tells us that he "developed a computer program to estimate the MNL model by maximum likelihood, a non-trivial task in those days." The development began in 1965, when "a Berkeley graduate student asked me for suggestions on how she might analyze her thesis data on freeway routing choices by the California Department of Highways. She completed her thesis before the program was working. However, I was eventually able to use the model to analyze her data (McFadden, 1968, 1976).''

Today, the fitting *is* a trivial task: Stata has a standalone `clogit` version; as readers will see in the bottom right portion of Figure 3, the version in `R` is simply a wrapper for a call to the function in the `survival` package that fits the Cox proportional hazards model. It takes advantage of the fact that the likelihood contribution from each riskset is of the same form as in conditional logistic regression (in Figure 2, each column of 4/5 days can be regarded as a 'riskset' (or a 'matched set' in the parlance of the economists). However, the calculations behind the fitting of the conditional logistic regression model are somewhat of a black box. The next section provides heuristics intended to make the fitting by Maximum Likelihood more transparent -- and the planning of case-crossover studies more intuitive.

The ML procedure for multinomial/conditional logistic regression, from first principles

The Method of Least Squares seeks the parameter value that minimizes the sum/average of the squared distances between the observed and fitted responses. Thus, since the quantity being 'optimized' uses the scale the responses are measured in, it is easily understood: if, for example, we fit a sine curve the pattern of temperatures over the year, the criterion involves the °C scale. Very differently, the Method of Maximum Likelihood seeks the parameter value that maximizes the sum/average of the logs of the probabilities of obtaining the data patterns that were observed. While the ML *principle* may be a natural one, the *scale* in which the criterion is measured is not so familiar. Nevertheless, as we will now see, the 'balancing equation' that must be satisfied/solved numerically *is* both natural and familiar.

To see why, we return to the data in column (5*) in Figure 2, where the temperatures on the 4 candidate days are 26, 24, 30 and 28 °C and, thus, the multinomial probabilities that the event occurred on the first, second third and fourth of these days are, respectively,

$$\frac{\{\, exp[26\beta],\ \ exp[24\beta],\ \ exp[30\beta],\ \ exp[28\beta]\,\}}{exp[26\beta]\ +\ exp[24\beta]\ +\ exp[30\beta]\ +\ exp[28\beta]}$$

The event occurred on the day when the temperature was 30 °C, and so the probability that it would have happened on that day rather than on one of the other three days is

$$\frac{exp[30\beta]}{exp[26\beta]\ +\ exp[24\beta]\ +\ exp[30\beta]\ +\ exp[28\beta]}$$

Thus, the log-likelihood contribution from this 'riskset' i.e., the log of this probability as a function of β, is

$$30\beta\ - \log\left(exp[26\beta]\ +\ exp[24\beta]\ +\ exp[30\beta]\ +\ exp[28\beta]\right)$$

The full log-likelihood is the sum, over the 10 risksets, of the riskset-specific contributions. To maximize it with respect to β, one finds the value at which its derivative equals zero. For the log-likelihood contribution from riskset (5*), the derivative with respect to β is

$$30\ - \frac{exp[26\beta]\ \times 26\ +\ exp[24\beta]\ \times 24 +\ exp[30\beta]\ \times 30 +\ exp[28\beta]\ \times\ 28}{exp[26\beta]\qquad +\ exp[24\beta]\qquad +\ exp[30\beta]\qquad +\ exp[28\beta]}.$$

Although it may seem formidable, the quantity after the minus sign is simply a weighted mean of the 4 temperatures, with weights given by the 4 exponentiated quantities. These weights are more manageable if we divide each of then by $exp[24\beta]$, so that the lowest temperature in the riskset receives a weight of 1, and so that the derivative (sometimes called the 'score') becomes

$$30\ - \frac{exp[2\beta]\ \times 26\ +\ 1\ \times 24 +\ exp[6\beta]\ \times 30 +\ exp[4\beta]\ \times\ 28}{exp[2\beta]\qquad +\ 1\qquad +\ exp[6\beta]\qquad +\ exp[4\beta]}.\ \ (1)$$

We can think of the quantity after the minus sign as *the "fitted" or "expected" value of the temperature on the day of the event*, and thus we can rewrite the equation in which the derivative is set to zero (often called the 'estimating equation') as the 'balancing equation'

Sum(*Observed* T on day of event) = Sum(*Fitted* T on day of event),

where the Sum is over the 10 risksets.

Today, unlike in 1965, the search for the ML estimate is easily carried out by trial and error in just a spreadsheet. As is shown at the right of Figure 2, the sum / mean of the observed temperatures on the 10 'event' days is 262 / 26.2 °C. If there were no linear relation with T, i.e., if $\beta = 0$, then the (null) fitted sum / mean would be 237.6 / 23.8 °C, and so we need to 'move up' β until the fitted sum / mean equals the observed value. This balance is achieved at $\beta = 0.261$, the same parameter estimate we saw earlier.

In Least Squares regression, the "y" residuals must balance each other. Perhaps not so surprisingly, since the conditioning *reverses* the x → y focus, in conditional logistic regression it is the residuals of the "x" values (the predictors in the regression) that must be balanced. Those familiar with fitting proportional hazards models may even recognize each difference between the observed and fitted temperature for the day of the event (shown in the last row of Figure 2) as a "Schoenfeld" residual. They will also have noted that there as many sets of residuals as there are predictors in the model, and that each set has as many residuals as there are risksets.

The Precision of the ML estimate of the exposure-response parameter

Before statistical packages were readily accessible, a first course in simple linear regression usually introduced the formula for the standard error of the fitted slope. Very often however, it was shown in a form that involved the fewest computational steps rather than for illumination, and so opportunities to gain some intuition as to what determines the precision were lost (Hanley 20xx). This loss is even greater in the case of model parameters fitted by ML, since the standard error is often model-based, and calculated only after the solution (often iterative) is reached. Thus, in the didactic spirit of this note, we will show how the standard error output by the `clogit` function is easily calculated from a mere spreadsheet. Since our conditional logistic regression model involves just 1 parameter, the 'matrix inversion' that is a feature of most regression fits takes the simple form of 1/I, where I is a scalar (1-dimensional) quantity. The reason for the choice of the letter I will become apparent later, and the 'I' quantity will play a central role in sample size projections.

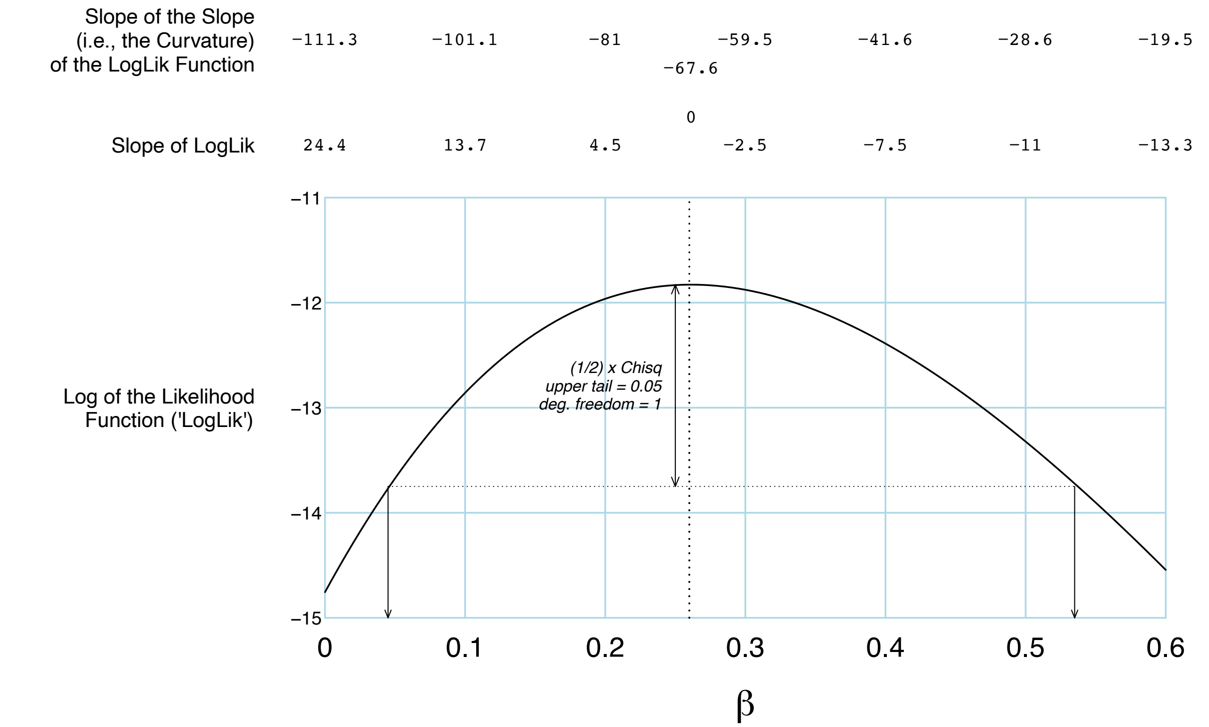


Figure 4:*Log-likelihood function for the parameter β of the exposure-response model, based on the data from the 10 matched sets in figure 2, together with its first and second derivatives computed at selected parameter values. The log-likelihood function reaches its maximum at β = 0.261, where its first derivative equals 0. The quantity 67.6 measures how curved the curve is at this ML value, and the square root of its reciprocal provides the Standard Error of the fitted β. The SE can then be used to form a Gaussian-based CI, or one can use the Likelihood ratio and the Ci-Square distribution to find the range of parameter values compatible with the data (limits are marked by the 2 arrows at 0.05 and 0.53).*

Before we introduce the formula-based approach that reveals where the precision (SE = 0.12) of the point estimate (0.261) comes from, we first use 'brute force' to numerically compute the standard error directly from the generic log-likelihood form. In other words, we rely *solely* on the log-likelihood function ('LogLik') plotted in the bottom of Figure 4. The ML estimate is the parameter value at which the first derivative (slope) of the log-likelihood function crosses from positive (at the left of the maximum) to negative (at the right), namely 0.261. Its variance is the reciprocal (inverse) of the (negative of the) second derivative of log-likelihood function evaluated at this same parameter value. This makes intuitive sense: the more concentrated (the sharper, or more curved) the curve is at its maximum, the narrower is the range of parameter values supported by the data. Moreover, as we go from left to right, the log-likelihood curve goes *from low to high to low*, so its slope (the first derivative) goes *from positive to negative*, and so its curvature (the second derivative) is *negative*. The more negative its is the tighter the log-likelihood and the more precise is the point estimate.

One can check manually/visually that the *first* derivative is 4.5 at $\beta = 0.2$ and-2.5 at $\beta = 0.3$, so the *second* derivative at $\beta = 0.25$ is approximately (-2.5 – 4.5)/0.1 or -70, and that its value of -67.6 at the ML value of $\beta = 0.261$ makes sense. R.A. Fisher, who developed the ML theory in the 1920s, called the -(-67.6) = 67.6 the '*Information*' (I) in the data concerning β, and showed that its *reciprocal* (i.e., 1/I = 1/67.6) can be taken as the *variance* of the β estimate, so that the Standard Error, the square root of the. variance, is

$$SE[\hat{\beta}] = (1/\text{Information})^{1/2} ,$$

or in this example,

$$SE[\hat{\beta}] = (1 / 67.6)^{1/2} = 0.1216,$$

in perfect agreement with the output from the `clogit` function.

Even though most textbooks begin their teaching of Maximum likelihood by defining the *Likelihood* as a *product* of probabilities, Fisher always began directly with the *log*-likelihood, so that it can be immediately written as a *sum* of the *individual* log-likelihood *contributions*, one from each 'datapoint'. Quite apart from making the sum a more manageable number, the log-version immediately emphasizes that each datapoint (or riskset in our example) adds to the information about the parameter of concern, that not all datapoints contribute equally, and that we can readily quantify in a technical sense exactly how much 'information' each one adds.

As we will now demonstrate, by working with this formal measure of information, and just taking the reciprocal of the combined information at the very end, the factors that determine the variance and the SE become very clear. So, instead of relying on the second derivative of the entire log-likelihood function as we did above, we will now show the specific *formula* that measures the 'information' contributed by each riskset, using as an example that contributed by riskset 5*. From equation (1) above giving the formula for the *first* derivative for the log-likelihood contribution, one can use the rules for derivatives to verify that the *second* derivative involves the same weights used in the weighed mean of the 4 temperatures, but that it is *the negative of the weighed means of the squared of the deviations of these 4 temperatures from that riskset-specific weighed mean*. The calculation of this weighted mean square is illustrated in Figure 2, where it is calculated under 3 scenarios: the null and ML values of β, and an intermediate value where $\beta = 0.1$. The 4 temperatures are 26, 24, 30 and 28, or (measured from their minimum, +2, 0, +6 and +4. Thus, at $\beta_{ML} = 0.261$, so that $\exp(\beta_{ML}) = 1.3$, the weights are $1.3^2 = 1.69$; 1; $1.3^6 = 4.79$; and $1.3^4 = 2.84$, so the weighted mean is mean is 28.2. The weighted mean of the squared deviations of the 4 temperatures from this 28.2 is 4.0. As such, it is the riskset with the second-smallest spread of temperatures, and it contributes the second smallest amount of information to the combined information of I = 67.6. The smallest contribution of the 10 risksets is the 3.6 from riskset (4) and the largest is the 10.9 from riskset (2). This ranking is the same as when the information is calculated at $\beta_{NULL} = 0$.

Readers may wonder why we do not refer to the weighed mean square deviation as a 'variance'. Technically it is, but since most associate a variance with a divisor that is one less than the numbers of objects, we prefer to use the more expressive term means square deviation. In his 1972 article, Cox refers to it as a "variance over the finite population of T's using an 'exponentially weighed' form of sampling." This fits with the principle that in a regression model, that x's are not treated as realizations of random variable whose variance is to be estimated (Hanley refs).

Fisher made a distinction between the *expected* information concerning β calculated using pre-study projections and the observed information calculated post study using the observed data. The latter is used to calculate the Standard Error for the β estimate, namely

$$SE[\,\hat{\beta}\,] = (1 / \mathrm{I})^{1/2} = (1 / [6.3 + 10.9 + \ldots\ 4.0 + \ldots + 6.8 + 8.6])^{1/2} = (1 / 67.6)^{1/2} = 0.1216.$$

The number of events (n.e.) to achieve a desired precision/power

For the precision, it is a matter of anticipating, pre-study, how large the information, I, concerning β will (or *is expected to*) be. Since I is a product of the number of events and the typical amount of information per riskset, the SE formula above can be inverted to give

$$number\ of\ events = \frac{1}{(SE_{desired})^2} \times \frac{1}{Amount\ of\ Information\ in\ a\ Typical\ Riskset},$$

i.e.,

$$\frac{1}{(SE_{desired})^2} \times \frac{1}{Weighted\ Mean\ Sq.Deviation\ of\ x\ values\ in\ a\ Typical\ Riskset}.$$

To plan for a given level of statistical power, the SE has to be envisioned under two scenarios, i.e., at the null, β_{null} (typically 0), and at the alternative, $\beta = \beta_{alt}$, so that they satisfy

$$Z_{\alpha/2} \times SE_{null} + Z_{\beta} \times SE_{alt} = \Delta,$$

where $\Delta = \beta_{alt} - \beta_{null}$ and where $Z_{\alpha/2}$ is typically 1.96 (for a 2-sided test with $\alpha = 0.05$), and Z_{β} is typically 0.84 (for 80% power).

Thus, if, say, we wished to have 80% power against an alternative of $\beta = 0.1$, we might use the data in Figure 2 as pilot data, and calculate that the typical information per riskset is 11.1 under the null and 10.1 under the alternative. Thus, the number of events, *n*, needs to satisfy the equation

$$\frac{1.96}{\sqrt{11.1 \times n}} + \frac{0.84}{\sqrt{10.1 \times n}} = 0.1,$$

or

$$number\ of\ events = \left(\left[\, 1.96 \div \sqrt{11.1} + 0.84 \div \sqrt{10.1}\ \right] \div 0.1 \right)^2.$$

One notices from figure 2 that the amount of information per riskset diminishes rapidly the further one departs from the null. So, a *conservative n* is obtained by using the non-null information for *both* SE's, so that, with these same error rates, the equation simplifies to

$$number\ of\ events = \left(\left[\, 2.8 \div \sqrt{NonNull\ Information\ in\ Typical\ Riskset}\ \right] \div \Delta \right)^2.$$

In our example, this comes out to 73 events if we use the more complex formula, or 78 if we use the conservative one.

Discussion (to be done)

What led us to write this article.

More general that the work of Künzli which is based only of triplets, not that intuitive, and formula may even be wrong.

Binary X is just a special case

Eg take say T >25 versus < 25.. redo analyses. Same SE formula.

Remarks on case crossover.. and on some of the studies cited below.

Not really case-crossover. Nothing personal.

Case crossover is for triggers, assuming that person is susceptible to start with.

Several of the studies have nothing personal in them.

If event was rain and motor vehicle collisions, or snow or ice, then its not a trigger.

Reveal what the events in Figure 1 are. Can you guess?

Our approach works for studies such as Redelmeier's one on weather

These studies yield rate ratios, NOT odds ratio, and not risk ratios.

Slippery slopes: multiplicity. P-value hacking using lags.

etc

References

A call for reporting the relevant exposure term in air pollution case-crossover studies

Nino Ku¨nzli, Christian Schindler

J Epidemiol Community Health 2005;59:527–530. doi: 10.1136/jech.2004.027391

Leah H. Schinasi, PhD, MSPH, Joan Rosen Bloch, CRNP, PhD, Steven Melly, MS, MA, Yuzhe Zhao, MS, Kari Moore, MS, and Anneclaire J. De Roos, PhD, MPH

High Ambient Temperature and Infant Mortality in Philadelphia, Pennsylvania: A Case–Crossover Study

February 2020, Vol 110, No. 2 AJPH

Schinasi et al.

Methods. We conducted a time-stratified case–crossover analysis of associations between ambient temperature and infant mortality in Philadelphia, Pennsylvania, during the warm months of 2000 through 2015. We used conditional logistic regression models to estimate associations of infant mortality with daily temperatures on the day of death (lag 0) and for averaging periods of 0 to 1 to 0 to 3 days before the day of death. We explored modification of associations by individual and census tract–level characteristics and by amounts of green space.

Results. Risk of infant mortality increased by 22.4% (95% confidence interval [CI] = 5.0%, 42.6%) for every 1°C increase in minimum daily temperature over 23.9°C on the day of death. We observed limited evidence of effect modification across strata of the covariates.

Sara Polcaro-Picheta,b, Tom Kosatskyc, Brian J. Potterd,e, Marianne Bilodeau-Bertrandb,

Environment International 126 (2019) 89–95 Contents lists available at ScienceDirect

Environment International

journal homepage: www.elsevier.com/locate/envint

Effects of cold temperature and snowfall on stroke mortality: A case- crossover analysis

Sara Polcaro-Pichet[a,b], Tom Kosatsky[c], Brian J. Potter[d,e], Marianne Bilodeau-Bertrand[b],

Nathalie Auger[a,b] ,[d] ,

[a] Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Quebec, Canada
[b] Institut national de santé publique du Québec, Montreal, Quebec, Canada
[c] National Collaborating Centre for Environmental Health, British Columbia Centre for Disease Control, Vancouver, Canada
[d] University of Montreal Hospital Research Centre, Montreal, Quebec, Canada
[e] Department of Cardiology, University of Montreal Hospital Center, Montreal, Quebec, Canada

T

Background: We sought to determine if cold temperature and snowfall are independently associated with stroke mortality, and whether effects differ between hemorrhagic and ischemic stroke.
Materials and methods: We conducted a case-crossover study of 13,201 stroke deaths utilizing weather records between the months of November and April for Quebec, Canada from 1981 to 2015. We compared exposure to cold temperature and snowfall with controls days when stroke

death did not occur. We computed odds ratios (OR) and 95% confidence intervals (CI) for the association of minimum temperature and duration of snowfall with stroke, adjusted for change in barometric pressure and relative humidity.

Results: The likelihood of mortality the day following exposure to cold temperature was elevated for hemorrhagic stroke in men, independent of snowfall. Relative to 0 °C, a temperature of − 20 °C was associated with 1.17 times the odds of hemorrhagic stroke death (95% CI 1.04–1.32). An independent effect of snowfall was als

------\\

Snowfall, Temperature, and the Risk of Death From Myocardial Infarction: A Case-Crossover Study

Wen Qi Gan∗, Sarah B. Henderson, Geoffrey Mckee, Weiran Yuchi, Kathleen E. McLean, Kris Y. Hong, Nathalie Auger, and Tom Kosatsky

AJE 2020 Vol. 189, No. 8

* Correspondence to Dr. Wen Qi Gan, Environmental Health Services, British Columbia Center for Disease Control, 655 West 12th Avenue, Vancouver, BC V5Z 4R4, Canada (e-mail: wenqi.gan@bccdc.ca).

The association between wind-related variables and stroke symptom onset: A case-crossover study on Jeju Island

Jayeun Kim, ... +3 ... , Jung-Kook Song Environmental Research • October 2016

Ambient temperature and risk of cardiovascular events at labor and delivery: A case-crossover study

Sandie Ha, ... +5 ... , Pauline Mendola Environmental Research • November 2017

Air pollution and humidity as triggering factors for stroke. Results of a 12-year analysis in the West Paris area

----

Rongbin Xu[a,b], Xiuqin Xiong[c], Michael J. Abramson[b], Shanshan Li[b,*], Yuming Guo[a,b,*]
Ambient temperature and intentional homicide: A multi-city case-crossover T study in the US

Environment International

Methods: We collected daily weather and crime data from 9 large US cities (Chicago, Detroit, Fort Worth, Kansas City, Los Angeles, Louisville, New York, Tucson and Virginia Beach) from 2007 to 2017. A time-stratified case- crossover design was used. The associations were quantified by conditional logistic regression with distributed lag models, adjusting for relative humidity, precipitation and effects of public holidays. City-specific odds ratios (OR) were used to calculate the attributable fractions in each city.

Results: Based on 19,523 intentional homicide cases, we found a linear temperature-homicide association. Every 5 °C increase in daily mean temperature was associated with a 9.5% [95% confidence interval (CI): 4.3–15.0%] and 8.8% (95% CI: 1.5–16.6%) increase in intentional homicide over lag 0–7 days in Chicago and New York, respectively. The association was not statistically significant in the other seven cities and seemed to be stronger for cases that happened during the hot season, at night (18:00–06:00) and on the street. During the study period, 8.7% (95%CI: 4.3–12.7%) and 7.1% (95% CI: 1.4–12.0%) intentional homicide cases could be attributed to temperatures above city-specific median temperatures, corresponding to 488 and 316 excess cases in Chicago and New York, respectively.

Author(s): Judith A McInnes, Muhammad Akram, Ewan M MacFarlane, Tessa Keegel, Malcolm R Sim and Peter Smith

Association between high ambient temperature and acute work-related injury: a case- crossover analysis using workers' compensation claims data

Source: Scandinavian Journal of Work, Environment & Health , January 2017, Vol. 43, No. 1 (January 2017), pp. 86-94

Methods A time-stratified case-crossover study design was used to examine the association between ambient temperatures and acute work-related injuries in Melbourne, Australia, 2002-2012, using workers' compensa tion claims to identify work-related injuries. The relationship was assessed for both daily maximum and daily minimum temperatures using conditional logistic regression.

Results Significant positive associations between temperature and acute work-related injury were seen for younger workers (<25 years), with the odds of injury increasing by 1% for each 1 °C increase in daily minimum temperature, and by 0.8% for each 1 °C increase in daily maximum temperature. Statistically significant asso ciations were also observed between daily maximum temperature and risk of injury for workers employed in the highest strength occupations and for male workers, and between daily minimum temperature and injury for all cases combined, female workers, workers aged 25-35 and >55 years, "light" and "limited" physical demand groups, and "in vehicle or cab" and "regulated indoor climate" workplace exposure groups.

Life-threatening alcohol-related traffic crashes in adverse weather: a double-matched case–control analysis from Canada

To cite: Redelmeier DA, Manzoor F. Life-threatening alcohol-related traffic crashes in adverse weather: a double- matched case–control analysis from Canada. BMJ Open 2019;**9**:e024415. doi:10.1136/ bmjopen-2018-024415

Donald A Redelmeier, Fizza Manzoor

mportance Drunk driving is a major cause of death in North America, yet physicians rarely counsel patients on the risks of drinking and driving.
Objective To test whether the risks of a life-threatening alcohol-related traffic crash were further accentuated by adverse weather.

Design Double matched case–control analysis of hospitalised patients.
setting Canada’s largest trauma centre between 1 January 1995 and 1 January 2015.

Participants Patients hospitalised due to a life- threatening alcohol-related traffic crash.
Exposure Relative risk of a crash associated with adverse weather estimated by evaluating the weather at the place and time of the crash (cases) compared with the weather at the same place and time a week earlier and a week later (controls).

results A total of 2088 patients were included, of whom the majority were drivers injured at night. Adverse weather prevailed among 312 alcohol-related crashes and was significantly more frequent compared with control circumstances. The relative risk of a life-threatening alcohol-related traffic crash was 19% higher during adverse weather compared with normal weather (95% CI: 5 to 35, p=0.006). The absolute increase in risk amounted to 43 additional crashes, extended to diverse groups of patients, applied during night-time and daytime, contributed to about 793 additional patient-days in hospital and was distinct from the risks for drivers who were negative for alcohol.

Conclusions Adverse weather was associated with an increased risk of a life-threatening alcohol-related traffic crash. An awareness of this risk might inform warnings to patients about traffic safety and counselling alternatives to drinking and driving.

Life-threatening motor vehicle crashes in bright sunlight

Donald A. Redelmeier, MD, FRCPC, MS(HSR), FACP[a,b,c,d,e,*], Sheharyar Raza, HBSc[a,b]

Abstract

Bright sunlight may create visual illusions that lead to driver error, including fallible distance judgment from aerial perspective. We tested whether the risk of a life-threatening motor vehicle crash was increased when driving in bright sunlight.

This longitudinal, case-only, paired-comparison analysis evaluated patients hospitalized because of a motor vehicle crash between January 1, 1995 and December 31, 2014. The relative risk of a crash associated with bright sunlight was estimated by evaluating the prevailing weather at the time and place of the crash compared with the weather at the same hour and location on control days a week earlier and a week later.

The majority of patients (n = 6962) were injured during daylight hours and bright sunlight was the most common weather condition at the time and place of the crash. The risk of a life-threatening crash was 16% higher during bright sunlight than normal weather (95% confidence interval: 9–24, P < 0.001). The increased risk was accentuated in the early afternoon, disappeared at night, extended to patients with different characteristics, involved crashes with diverse features, not apparent with cloudy weather, and contributed to about 5000 additional patient-days in hospital. The increased risk extended to patients with high crash severity as indicated by ambulance involvement, surgical procedures, length of hospital stay, intensive care unit admission, and patient mortality. The increased risk was not easily attributed to differences in alcohol consumption, driving distances, or anomalies of adverse weather.

Bright sunlight is associated with an increased risk of a life-threatening motor vehicle crash. An awareness of this risk might inform driver education, trauma staffing, and safety warnings to prevent a life-threatening motor vehicle crash.

Level of evidence: Epidemiologic Study, level III.
Abbreviations: None.

Biometrics & Biostatistics

Hanley and Moodie, J Biomet Biostat 2011, 2:5

http://dx.doi.org/10.4172/2155-6180.1000124

Sample Size, Precision and Power Calculations: A Unified Approach

James A Hanley* and Erica EM Moodie

http://www.medicine.mcgill.ca/epidemiology/hanley/Reprints/UniversalSampleSize.pdf

https://www.nobelprize.org/uploads/2018/06/mcfadden-lecture.pdf

J. R. Statist. Soc. A (2008) **171**, *Part* 2, *pp.* 375–394

Applying discrete choice models to predict Academy Award winners

Iain Pardoe

University of Oregon, Eugene, USA

and Dean K. Simonton

Biometrics 59, 1–8

March 2003

PRESIDENTIAL ADDRESS: XXI International Biometric Conference, Freiburg, Germany, July 2002

Are Statistical Contributions to Medicine Undervalued? Norman E. Breslow

https://www.nobelprize.org/uploads/2018/06/mcfadden-lecture.pdf

----

OXFORD UNIVERSITY PRESS | International Journal of Epidemiology

Planning and understanding sample sizes for case-crossover studies of environmental exposures

Journal:	*International Journal of Epidemiology*
Manuscript ID	Draft
Manuscript Type:	Education Corner
Date Submitted by the Author:	n/a
Complete List of Authors:	Hanley, JA; McGill University Montreal, Department of Epidemiology, Biostatistics, and Occupational Health Weichenthal, Scott; McGill University
Key Words:	Parameter estimation, Conditional Logistic Regression, Exposure Variation, matched sets, risksets

2022.03.21

For: Education Corner, IJE

Planning and understanding sample sizes for case-crossover studies of environmental exposures

James A. Hanley[1*]and Scott Weichenthal[1]

[1]Department of Epidemiology, Biostatistics, and Occupational Health,
McGill University, Montreal, Canada

*Corresponding author. Department of Epidemiology, Biostatistics, and Occupational Health. McGill University, Montreal, H3A 1G1, Canada. E-mail: James.Hanley@mcgill.ca

Abstract

Time-stratified case-crossover studies are often used to quantify the relationship between the rates of acute health events and levels of environmental exposures such as heat or air pollution. Especially when exposures are to be measured on a continuous scale, few sample-size planning tools are available to anticipate the statistical precision of the resulting effect estimate, or to appreciate the study design aspects that influence statistical power. We provide formulae that can be used to plan the sizes of time-stratified case-crossover studies with exposures measured on either a categorical or continuous scale. We explain where the formulae come from using a small hand-worked example. We illustrate the Maximum Likelihood (ML) calculations involved in estimating parameters from the relevant conditional logistic regression models. The expected amount of statistical 'information' that each matched set contributes to the ML parameter estimate is emphasized. The precision of the estimated regression coefficient in time-stratified case-crossover studies depends on both the number of cases studied and the variation in the exposure values within a typical matched set (as measured by the Mean Squared Deviation). Importantly, the within-matched-set variation in continuous exposures will often be much less than the variance of exposures observed over the duration of the study period (e.g., daily variations in outdoor temperatures during 2022 compared with variation of daily temperatures for all Fridays in July 2022). Investigators conducting such studies should pay close attention to the expected within-set variation in exposures to ensure that an adequate number of cases is identified. The same considerations apply to case-crossover studies of non-environmental exposures.

Keywords:

Parameter estimation; Conditional Logistic Regression; Exposure Variation; risksets; matched sets

Word count: 2008

Key Messages

- Simple formulae can be used to plan the sizes of time-stratified case-crossover studies with exposures measured on either a categorical or continuous scale.

- The precision of the estimated regression coefficient in time-stratified case-crossover studies depends on both the number of cases studied and the variation in the exposure values within a typical matched set (as measured by the Mean Squared Deviation).

- Investigators should pay close attention to the expected within-set variation in exposures to ensure that an adequate number of cases is identified.

- The same statistical considerations apply to case-crossover studies of non-environmental exposures.

- The basis for the formulae can be understood by working through a small hand-worked example.

Introduction

Time-stratified case-crossover studies are commonly used to estimate the acute health impacts of environmental exposures such as heat or air pollution.[1] This design begins by identifying cases (e.g., people admitted to hospital for a myocardial infarction). For each case, exposure data are obtained for the day of the event (or a time-period immediately prior to it), and for other days with similar characteristics (i.e., the same day of the week, month, and year). For example, if a person experienced the event on a Friday in May 2021, the exposure data pertaining to this case might be assembled for all Fridays in May 2021. We will refer to this *set of days* (which *includes* the day the event occurred) as a *matched-set*. The matched-set is the conceptual counterpart of the risk-set in a survival analysis, or in a standard incidence-density-based matched case-control study. Typically, conditional logistic regression models are used to fit the event rate as a function of the exposure levels.

Despite the common use of the case-crossover design in environmental epidemiology, guidance on factors that determine sample size and statistical power in these studies is not readily available. To address this gap, we provide formulae that can be used to calculate these quantities and illustrate the theory behind these equations using a simple hand-worked example. We emphasize that the amount of statistical 'information' each case contributes to the parameter estimate can be quantified by the typical Mean Square Deviation (MSD) within a typical matched set. If this MSD is expected to be small, a larger number of cases must be included. Although our focus is on continuous exposure measures typical of environmental epidemiology, the same statistical principles apply to binary or categorical exposures, and to non-environmental exposures. Further technical details are provided in the online supplement.

Preliminaries

The parameter of interest

For concreteness, we begin with an example where some aspect (e.g., mean, maximum) of the daily temperature is the exposure of interest. We simply refer to this as *'T'*. Suppose the model we will use to relate the event rate (i.e., the expected number of events per day) to *T*, is

$$\lambda(T) = \lambda_0 \times \exp[\,\beta\,(T - T_0\,)\,]\,, \tag{1}$$

where λ_0 refers to the event rate at some reference temperature, T_0 and the shorthand ‘exp’ stands for ‘the exponentiated value of.’ Thus, $\exp[\,\beta\,(T - T_0\,)\,]$ denotes the ratio of the event rate at temperature T to the rate at the reference temperature; if T is measured in degrees Celsius, then β, the parameter to be fitted to (estimated from) the data, refers to the log of the ratio of the event rates at temperatures that are 1° Celsius apart.

Importantly, we will divide our presentation into two scenarios, which we *arbitrarily* divide into ‘weaker’ and ‘stronger’ exposure-response relationships. By ‘weaker’ we mean a coefficient β such that, over the T range in a typical matched set, the rate at the upper end is less than 1.1 times the reference rate (of 1) at the lower end. As we will see below, the sample size calculations in the ‘weak’ scenario are considerably simpler.

The spread of the exposure data in a typical matched set

Suppose that for a typical matched set of 4 days in a case-crossover study, the exposures (values of T) for the 4 days in the set are: 21°C, 23°C, 25°C and 27°C. Of course, temperatures will not usually be so rounded or so regular: these 4 values were selected to make for convenient calculations. The mean of these values is 24 and the mean squared deviation (MSD) from 24 is 5 ($[(24\text{-}21)^2 + (24\text{-}23)^2 + (24\text{-}25)^2 + (24\text{-}27)^2] / 4 = (9+1+1+9)/4 = 20/4 = 5$). Note that the MSD of 5 is the same as if we had recoded the 4 *T*s as 0, 2 4 and 6°C above the minimum in the set. It is important to note that this MSD is smaller than the *sample variance* of the 4 values. The sum of the 4 squared deviations is divided by 4, not 3, since we are not *estimating* a population variance, but rather measuring how spread out the 4 *T*s are.

With these preliminaries, we first address the number of cases (and thus the number of matched sets) to ensure that the regression coefficient β will be estimated with a specified level of precision. Since the 'weaker-relationship' scenario is more common, we begin with it; as it happens, the calculations in this context are also simpler.

Weaker-relationship scenario

Number of cases to ensure a desired precision

Suppose that we set the precision with which the regression coefficient β will be estimated by specifying that its 95% margin of error (ME) will not exceed some specified amount. This implies that its Standard Error (SE) will not exceed 1/2 (technically 1/1.96) of this ME. The number (n) of events required to achieve this SE is given by the formula

$$n = \frac{1}{(SE_{desired})^2} \times \frac{1}{Mean\ Sq.\ Deviation\ of\ exposure\ values\ in\ a\ Typical\ Matched\ Set}.$$

It makes sense that the MSD is in the *denominator* of the formula, just as it is in the expression for the variance of a slope in a simple regression, where the narrower/wider the spread of the x's the more/less stable will be the fitted slope.[2] As an example, suppose the T's in a typical matched are expected to have a MSD of 5. Suppose that, relative to the reference T_0, the anticipated rate ratio at $T_0 + 1$ is 1.05, so that $\beta = \ln(1.05) = 0.049$. Suppose we wish the SE for the fitted β to be no larger than 0.02 (or that the margin of error not exceed 0.04). Then, to achieve this, we would need to study

$$n = \frac{1}{(0.02)^2} \times \frac{1}{5} = 500 \text{ events.}$$

If we we wish the SE to be no larger than 0.01 (or the margin of error to not exceed 0.02), we would need to study $n = (1/0.01)^2 / 5 = 2{,}000$ events, i.e., it takes 4 time as many cases to cut the margin of error in 2.

Number of cases to ensure a specified power

The sample size to guarantee a pre-specified power (of say 80%) is larger than when (in the absence of null hypothesis testing) precision is the only concern. The larger requirement stems from the added insistence on an 80% chance that (under the alternative) the point estimate exceeds the criterion for a ‘statistically significant’ result. Typically $Z_{\alpha/2} = 1.96$ for a 2-sided test with $\alpha = 0.05$, and $Z_\beta = 0.84$ for 80% power. Thus, if Δ is the difference between the alternative and null values of β, the required number of events n is

$$n = \frac{(1.96 + 0.84)^2}{\Delta^2} \times \frac{1}{Mean\ Squared\ Deviation\ of\ Exposures\ in\ Typical\ Matched\ Set}$$

In our example, if the alternative (to the null $\beta = 0$) is $\beta = 0.049$, and the anticipated Mean Squared Deviation of the T's in a typical matched set is 5, then we require

$$n = \frac{2.8^2}{0.049^2} \times \frac{1}{5} = 653 \text{ events.}$$

The ‘anatomy’ of this formula is similar to that of equation (5) in a not-well known but quite instructive 1985 article.[3]

Stronger-relationship scenario

Number of cases to ensure a desired precision

For a desired degree of precision, the formula has the same form as the earlier one, except that each MSW is now a weighted MSW, and thus narrower that the un-weighed one. (Why it is smaller is explained in the Supplement). However, since the β is larger than in our initial example, the required number of events may be smaller. To make these aspects concrete, suppose that, relative to the reference T_0, the anticipated rate ratio at $T_0 + 1$ is 1.2, so that β = ln(1.2) = 0.18. Suppose we wish the SE for the fitted β to be no larger than 0.05 (or that the margin of error not exceed 0.10). To get a sense of a typical weighted MSD, we might treat the data in Figure 1A as if they came from a pilot study. Consider first the relatively narrow spread of *T*s in matched set 5, namely 24, 26, 28 and 30 (or 0, +2, +4 and +6 above the minimum in the matched set). With a Rate Ratio of 1.2 per degree C, the weights are 1.2^0 =1, 1.2^2=1.44, 1.2^4= 2.07 and 1.2^6= 2.99, so that the weighted MSW is 4.5 (around a weighted mean of 27.9 C). At the other (more favourable) extreme, consider the more spread out T's of 18.5, 21.5, 25 and 30 in matched set 2 (or 0, +3, +6.5 and +11.5 above the minimum in the set): the weighted MSD in this set is 14.9. To be *conservative,* we might take the typical weighted MSD to be on the '*smaller'* side, say 4.5. Under this almost-worst case scenario, to achieve the SE of 0.05, we would need to study $\frac{1}{0.05^2} \times \frac{1}{4.5} = 89$ events.

Number of cases to ensure a specified power

To plan for a given level of statistical power, the SE of $\hat{\beta}$ has to be envisioned under *two* scenarios, i.e., at the null, β_{null} (typically 0), and at the alternative, $\beta = \beta_{alt}$, so that they satisfy

$$Z_{\alpha/2} \times SE_{null} + Z_{\beta} \times SE_{alt} = \Delta,$$

where $\Delta = \beta_{alt} - \beta_{null}$.

Thus, if, say, we wished to have 80% power against an alternative of $\beta = 0.18$, we might use the data in Figure 1A as pilot data, and calculate (conservatively) that the typical weighted MSD per riskset will be 5 under the null and 4.5 under the alternative. Thus, the number of events, *n*, needs to satisfy the equation

$$\frac{1.96}{\sqrt{\text{MSW}_{null} \times n}} + \frac{0.84}{\sqrt{\text{MSW}_{alt} \times n}} = \Delta.$$

Thus, with our anticipated $\text{MSW}_{null} = 5$, and $\text{MSW}_{alt} = 4.5$, the required number of events would be

$$n = \left(\left[\, 1.96 \div \sqrt{5.0} + 0.84 \div \sqrt{4.5}\ \right]\ \div 0.18 \right)^2 = 50\ events.$$

One notices from Figure 1B that the amount of information per matched set diminishes rapidly the further β departs from the null. So, a *conservative n* is obtained by using the *non-null* information for *both* SE's. With these same error rates, and noting that 1.96+0.84 = 2.8, the equation simplifies to

$$\text{number of events} = \left(\left[\, 2.8 \div \sqrt{\text{NonNull MSD in Typical Riskset}}\ \right]\ \div \Delta \right)^2.$$

If we want an easier to remember (and again slightly conservative) formula, we can round 2.8^2 up to 8, to obtain

$$\text{number of events} = (8\ \div \text{NonNull MSD in Typical Riskset})\ \times (1/\Delta\,)^2\,.$$

In our example, with $(1/0.18)^2$ rounded up to 31, this comes out to (8/4.5) × 31 = 55 events.

Discussion

Intuitively, greater variation in the exposure makes it easier to detect/measure an exposure-response relationship. Since time-stratified case-crossover studies make comparisons *within* each time-matched set, the precision/power depends on the *within-matched-set* variation, and not on the *overall* variation in the exposure.[4] This ‘local’ variation can be much smaller: for example, in Figure 1A, while the MSD of the 44 temperatures around the overall mean of 23.8 is 19.4 C^2, the typical within-matched-set MSD is only11.1 C^2. Thus, investigators need to pay attention to the expected within-set variation in exposures to ensure that an adequate number of cases is identified.

Unless the exposure-response relationship is quite strong, ‘local’ MSDs calculated at the null will suffice for planning purposes, since a sample size exercise is merely a rough projection of the likely precision/power. As the authors of a classic textbook[5] cautioned “There is usually little point in introducing fine detail into what are essentially rather crude calculations.”

Investigators should not base them on implausibly large values of Δ, or think that any one study will settle the matter. Instead, they should consider how much information their study will contribute to a future meta-analysis. A former colleague of ours likened the question to how much to give when the collection plate is passed around in a house of worship: it is the *total* collected that matters in the end; in most such places, there is no ‘requirement’ for the size of an individual contribution.[6]

Lastly, as we explain in the Online Supplement, rather than present separate formulae for exposures measured on continuous and all-or-none exposures, we urge investigators to use the common principles involved. And, of course, the same considerations apply to case-crossover studies of non-environmental exposures.

References

1 Janes H, Sheppard L, Lumley T. Case–Crossover Analyses of Air Pollution Exposure Data: Referent Selection Strategies and Their Implications for Bias. *Epidemiology* 2005;16:717-726.

2 Hanley JA. Simple and multiple linear regression: sample size considerations. *Journal of Clinical Epidemiology* 2016;79:112-119.

3 McKeown-Eyssen GE, Thomas DC. Sample size determination in case-control studies: the influence of the distribution of exposure. *J Chronic Disease* 1985;38:559-568.

4 Künzli N, Schindler C. A call for reporting the relevant exposure term in air pollution case-crossover studies . *J Epidemiol Community Health* 2005;59:527-530.

5 Breslow NE, Day NE. Design Considerations. Chapter 7 in Statistical Methods in Cancer Research Volume II: The Design and Analysis of Cohort Studies 1987. IARC Scientific Publication No. 82.

6 Hernán MA. Causal analyses of existing databases: no power calculations required*. J Clinical Epidemiology* 2021: Aug 27;S0895-4356(21)00273-0. doi: 10.1016/j.jclinepi.2021.08.028.Online ahead of print.

APPENDIX / ONLINE SUPPLEMENT

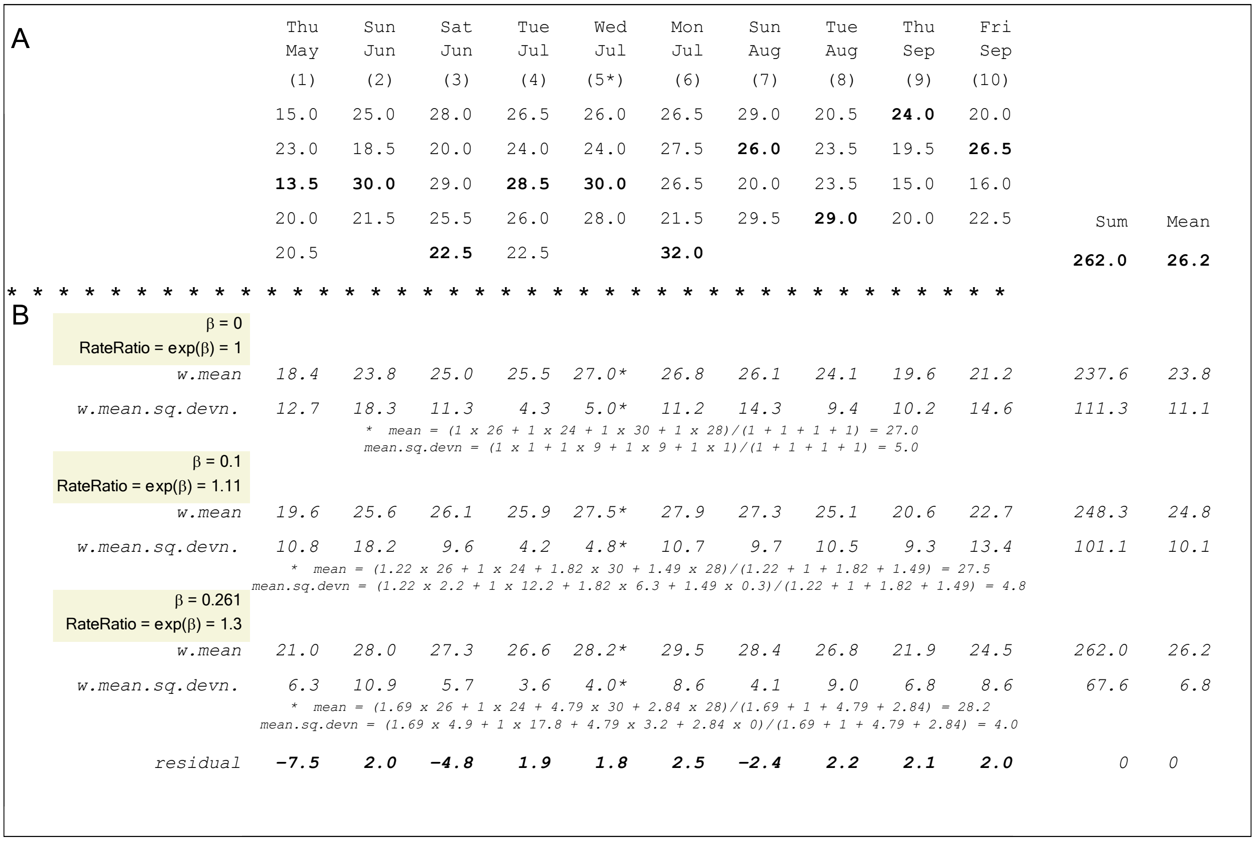
A WHERE DO THE FORMULAE COME FROM?

Sample size calculations are *pre-study* calculations that depend on the data-analysis method that will ultimately be used (i.e., post data-collection) Thus, to understand them, it is best to go through an actual data-analysis exercise, and to *anticipate* the results of the model-fitting. To this end we begin with a small dataset and to see close-up what aspects of the data determine the standard errors that emerge during the parameter-fitting. To keep the dataset small but real, we studied tornadoes, where the relationship between *T* and their rate is strong enough to 'see' in a study of just 10 cases. [To have the study design mimic a study of human events, we retain the matching on day-of-the-week and month]

Part A of Figure 1 shows the *T*'s for each of 10 matched sets generated by the 10 tornadoes that occurred in the southern portion of a Canadian province during one selected year. Without loss of generality, we consider the temperature (*T*) on a specified day, rather than a lagged version of *T*, as the determinant of the expected event rate for that day. We limit ourselves to the same multiplicative model for event rates shown in equation (1) above.

The SE of the β fitted by conditional logistic regression to the dataset in Figure 1

The average of the 10 T's on the 10 'event' days was 26.2˚C, whereas, as is shown in the first row of part B, the average of the 10 column-specific averages was only 23.8˚C. This indicates that the sign of the fitted gradient of the event rates over T will be positive.



A

	Thu May (1)	Sun Jun (2)	Sat Jun (3)	Tue Jul (4)	Wed Jul (5*)	Mon Jul (6)	Sun Aug (7)	Tue Aug (8)	Thu Sep (9)	Fri Sep (10)	Sum	Mean
	15.0	25.0	28.0	26.5	26.0	26.5	29.0	20.5	**24.0**	20.0		
	23.0	18.5	20.0	24.0	24.0	27.5	**26.0**	23.5	19.5	**26.5**		
	13.5	**30.0**	29.0	**28.5**	**30.0**	26.5	20.0	23.5	15.0	16.0		
	20.0	21.5	25.5	26.0	28.0	21.5	29.5	**29.0**	20.0	22.5		
	20.5		**22.5**	22.5		**32.0**					**262.0**	**26.2**

* *

B

	(1)	(2)	(3)	(4)	(5*)	(6)	(7)	(8)	(9)	(10)	Sum	Mean
β = 0 RateRatio = exp(β) = 1												
w.mean	*18.4*	*23.8*	*25.0*	*25.5*	*27.0**	*26.8*	*26.1*	*24.1*	*19.6*	*21.2*	*237.6*	*23.8*
w.mean.sq.devn.	*12.7*	*18.3*	*11.3*	*4.3*	*5.0**	*11.2*	*14.3*	*9.4*	*10.2*	*14.6*	*111.3*	*11.1*
β = 0.1 RateRatio = exp(β) = 1.11												
w.mean	*19.6*	*25.6*	*26.1*	*25.9*	*27.5**	*27.9*	*27.3*	*25.1*	*20.6*	*22.7*	*248.3*	*24.8*
w.mean.sq.devn.	*10.8*	*18.2*	*9.6*	*4.2*	*4.8**	*10.7*	*9.7*	*10.5*	*9.3*	*13.4*	*101.1*	*10.1*
β = 0.261 RateRatio = exp(β) = 1.3												
w.mean	*21.0*	*28.0*	*27.3*	*26.6*	*28.2**	*29.5*	*28.4*	*26.8*	*21.9*	*24.5*	*262.0*	*26.2*
w.mean.sq.devn.	*6.3*	*10.9*	*5.7*	*3.6*	*4.0**	*8.6*	*4.1*	*9.0*	*6.8*	*8.6*	*67.6*	*6.8*
residual	***−7.5***	***2.0***	***−4.8***	***1.9***	***1.8***	***2.5***	***−2.4***	***2.2***	***2.1***	***2.0***	*0*	*0*

For β = 0:

```
*  mean = (1 x 26 + 1 x 24 + 1 x 30 + 1 x 28)/(1 + 1 + 1 + 1) = 27.0
mean.sq.devn = (1 x 1 + 1 x 9 + 1 x 9 + 1 x 1)/(1 + 1 + 1 + 1) = 5.0
```

For β = 0.1:

```
*  mean = (1.22 x 26 + 1 x 24 + 1.82 x 30 + 1.49 x 28)/(1.22 + 1 + 1.82 + 1.49) = 27.5
mean.sq.devn = (1.22 x 2.2 + 1 x 12.2 + 1.82 x 6.3 + 1.49 x 0.3)/(1.22 + 1 + 1.82 + 1.49) = 4.8
```

For β = 0.261:

```
*  mean = (1.69 x 26 + 1 x 24 + 4.79 x 30 + 2.84 x 28)/(1.69 + 1 + 4.79 + 2.84) = 28.2
mean.sq.devn = (1.69 x 4.9 + 1 x 17.8 + 4.79 x 3.2 + 2.84 x 0)/(1.69 + 1 + 4.79 + 2.84) = 4.0
```

***Figure 1**: A: the temperatures,(T, in ° Celsius) for the day-of-week-that-month 'strata' or 'matched sets' containing the 10 events that occurred in a selected year. The temperature on the **day of the event** is indicated in **bold**. The asterisk in column 5 indicates that the ML calculations are presented in full for this selected column.*

B: the calculations used in the pursuit of the Maximum Likelihood (ML) estimate of β, starting with the null value. Each `mean` *is a weighted average of the temperatures $T_1, T_2, \ldots T_{4/5}$ within a stratum, with weights $\exp\,[\beta T_1]$, $\exp\,[\beta T_2]$, … or, equivalently, as shown, with re-scaled weights $\exp\,[\beta T'_1]$,$\exp\,[\beta T'_2]$, … where $T'_1, T'_2, \ldots T'_{4/5}$ are measured relative to the minimum T in the stratum, Thus, the minimum temperature in the stratum has a weight of 1.*

Each `mean.sq.devn` *is a weighted average of the squared deviations of* $T_1, T_2, \ldots T_{4/5}$ *from* `mean`*, using these same weights. (For the calculations involving*$\beta = 0$*, all values in the matched set receive the same weight). The detailed calculations are shown for the selected column (5*). The sum/mean at the right is the sum/mean over the 10 instances/cases. The ML iterations continue until the sum/mean of the 10 fitted/weighted means equals (balances) the sum/mean of the 10 (observed) temperatures on the days the events occurred.*

As we show in section B, the Maximum Likelihood value of β (0.261) is found by starting at β = 0 and proceeding, by a directed search, until one reaches a β value for which the sum of the *T*'s on the 10 days when the event occurred (the '*observed*' Ts) *equals* the sum of the '*fitted*' *T*'s on these 10 days. More important is the formula for its standard error, $SE[\hat{\beta}]$ = 0.1216. As we explain in the supplement, the SE is found by summing the MSD's in the 10 matched sets to arrive at a total of 67.6, and taking the square root of the reciprocal of this, i.e., $(1 / 67.6)^{1/2}$ = 0.1216. Note, however, that the 10 matched-set-specific MSD's are not the 12.7, 18.3, … 14.6 in the first set of calculations in Figure 1B. Since β = 0, these 'initial' MSD's are calculated by weighing the T's within the set equally; they sum to 111.3. The MSDs that sum to 67.6 were calculated as *weighted* MSD's, where the weights for the T's in each matched set are the rate ratios implied by the value of β and the T's in the set. The calculation of the weighted MSD is illustrated for matched set 5. In it, when β *= 0.261,* the weights for the 4 *T*s of 24, 26, 28 and 30 (or *T*s of 0, +2, +4 and +6 above the minimum in the set) are exp(0 × 0.261) = <u>1</u>, exp(2 × 0.261) = <u>1.69</u>, exp(4 × 0.261) = <u>2.84</u> and exp(6 × 0.261) = <u>4.79</u> respectively. The MSD for matched set 5 was 4.0; the MSDs for the 9 other matched sets ranged from 3.6 to 10.9, and the typical MSD was 6.8.

It can be seen from Figure 1B that the further β is from 0, the smaller is the typical weighted MSD, and thus the larger is the SE of the fitted β: Whereas the SE is $(1 / 67.6)^{1/2}$ = 0.12 at the ML value of β = 0.261, it is $(1 / 111.3)^{1/2}$ = 0.09 at the null value of β. Thus, when β is further from zero, the required sample sizes will be larger than those illustrated in the earlier sections.

B MAXIMUM LIKELIHOOD ESTIMATION DEMYSTIFIED

Multinomial probabilities: the possible days an event could have occurred

As Chapters 13 and 15 of Clayton and Hills[1] show, and as Armstrong[2] re-iterates, the rate ratio in a person-time analysis of a binary exposure can sometimes be estimated by treating the total number of events within the stratum as a fixed quantity rather than the random variable that it is. In the examples addressed in these chapters, how the events distribute themselves within the ‘exposed’ and ‘unexposed’ person-time can be described by a *binomial* random variable, in which the number of events serves as the ‘*n*’ and the probability parameter is a function of the amounts of person time and the rate ratio. Parameter estimation is usually via Maximum Likelihood (ML). In *our* context, where the possible event days are days within a matched set, how the events distribute themselves over the possible days can be described by a *multinomial* random variable, in which the number of events (typically 1 per matched set) serves as the '*n*' and the multinomial probabilty parameters are a function of the temperatures and the rate ratios. For example, in column (5*) in Figure 1 in the main text, the temperatures on the 4 candidate days are 26, 24, 30 and 28 ˚C. Thus, *given* that an event occurred on one of these days (i.e., *conditional* on the event having occurred within the stratum), the multinomial probabilities that it occurred on the first, second, third or fourth of these days are, respectively,

$$\frac{\{\; exp[26\beta],\;\; exp[24\beta],\;\; exp[30\beta],\;\; exp[28\beta]\;\}}{exp[26\beta]\; +\; exp[24\beta]\; +\; exp[30\beta]\; +\; exp[28\beta]}\,.$$

These probabilities have the same structure as the probabilities that each of the nominees will win the Oscar[4-6] or the economic choices made by a consumer.[7-8]

The ML procedure for multinomial/conditional logistic regression, from first principles

The Method of Least Squares seeks the parameter value that minimizes the sum/average of the squared distances between the observed and fitted responses (the 'y's). Thus, since the quantity being ‘optimized’ uses the scale the responses are measured in, it is easily understood: if, for example, we fit a sine curve to the pattern of temperatures over the year, the criterion involves discrepancies in the ˚C scale. Very differently, *the Method of Maximum Likelihood seeks the parameter value that maximizes the sum/average of the logs of the probabilities of obtaining the data patterns that were observed.* While the ML *principle* may be a natural one, the *scale* in which the criterion is measured is not so familiar. Nevertheless, as we will now see, the ‘balancing equation’ that must be satisfied/solved numerically is quite natural, even if not always emphasized.

To see why, we return to the data in column/stratum (5) in Figure 1 in the main text, where the temperatures on the 4 candidate days are 26, 24, 30 and 28 ˚C and, thus, the multinomial probabilities that the event occurred on the first, second, third or fourth of these days are, respectively,

$$\frac{\{\, exp[26\beta],\ \ exp[24\beta],\ \ exp[30\beta],\ \ exp[28\beta]\,\}}{exp[26\beta]\ +\ exp[24\beta]\ +\ exp[30\beta]\ +\ exp[28\beta]}$$

The event occurred on the day when the temperature was 30 ˚C, and so the probability that it would have happened *on that day (rather than on one of the other three days)* is

$$\frac{exp[30\beta]}{exp[26\beta]\ +\ exp[24\beta]\ +\ exp[30\beta]\ +\ exp[28\beta]}$$

Thus, the log-likelihood contribution from this matched-set, i.e., the log of this probability as a function of β, is

$$30\beta\ - \log\left(exp[26\beta]\ +\ exp[24\beta]\ +\ exp[30\beta]\ +\ exp[28\beta]\right)$$

The full log-likelihood is the sum, over the 10 matched sets, of the set-specific contributions. To maximize it with respect to β, one finds the value at which its derivative equals zero. For the log-likelihood contribution from matched set (5*), the derivative with respect to β is

$$30\ - \frac{exp[26\beta]\ \times 26\ +\ exp[24\beta]\ \times 24 +\ exp[30\beta]\ \times 30 +\ exp[28\beta]\ \times\ 28}{exp[26\beta]\qquad +\ exp[24\beta]\qquad +\ exp[30\beta]\qquad +\ exp[28\beta]}.$$

Although it may seem formidable, the quantity to the right of the minus sign is simply a *weighted mean of the 4 temperatures*, with weights given by the 4 exponentiated quantities. These weights are more manageable if we divide each of them by $exp[24\beta]$, so that the *lowest temperature in the matched set receives a weight of 1*, and so that the derivative (sometimes called the ‘score’) becomes

$$30\ - \frac{exp[2\beta]\ \times 26\ +\ 1\ \times 24 +\ exp[6\beta]\ \times 30 +\ exp[4\beta]\ \times\ 28}{exp[2\beta]\qquad +\ 1\qquad +\ exp[6\beta]\qquad +\ exp[4\beta]}.\ (A1)$$

We can think of the quantity after the minus sign as the "fitted" or "expected" value of the temperature on the day of the event, and thus we can rewrite the equation in which the derivative is set to zero (often called the 'estimating equation') as the 'balancing equation'

$$\text{Sum(Observed } T \text{ on day of event)} = \text{Sum(Fitted } T \text{ on day of event)},$$

where the Sum is over the 10 matched sets.

Today, unlike when this model was first fitted in the mid 1960s, the search for the ML estimate can be easily carried out by trial and error using just a spreadsheet. As is shown at the right of Figure 1, the sum / mean of the observed temperatures on the 10 'event' days is 262 / 26.2 ˚C. If there were no linear relation with T, i.e., if $\beta = 0$, then the (null) fitted sum / mean would be 237.6 / 23.8 ˚C. Since 26.2 is larger than expected, we need to 'move up' β until the fitted sum / mean equals the observed value. As can see seen in Figure 1, this 'balance' is achieved at $\beta = 0.261$.

In Least Squares regression, the "y" residuals must balance each other. Perhaps not so surprisingly, since *conditioning* reverses the $x \rightarrow y$ focus, in conditional logistic regression it is the residuals of the "x" values (the predictors in the regression) that must be balanced. Those familiar with fitting proportional hazards models may even recognize each difference between the observed and fitted temperature for the day of the event (shown in the last row of Figure 1) as a “Schoenfeld” residual.

The Precision of the ML estimate of the exposure-response parameter

Before statistical packages were readily accessible, a first course in simple linear regression usually introduced the closed-form formula for the standard error of the fitted slope. Very often however, it was shown in a form that involved the fewest computational steps rather than for illumination, and so opportunities to gain some intuition as to what determines the precision were lost.[9] This lack of transparency is even greater in the case of parameters fitted by ML, since the standard error is model-based, and calculated only after the solution (often iterative) is reached. Thus, in the didactic spirit of this note, we will show how the standard error output by a conditional logistic regression routine is easily calculated from a mere spreadsheet. Since our conditional logistic regression model involves just 1 parameter, the ‘matrix inversion’ that is a feature of most regression fits takes the simple form of $1/I$, where I is a scalar (1-dimensional) quantity. The reason for the choice of the letter I will become apparent later, and the ‘I’ quantity will play a central role in sample size projections.

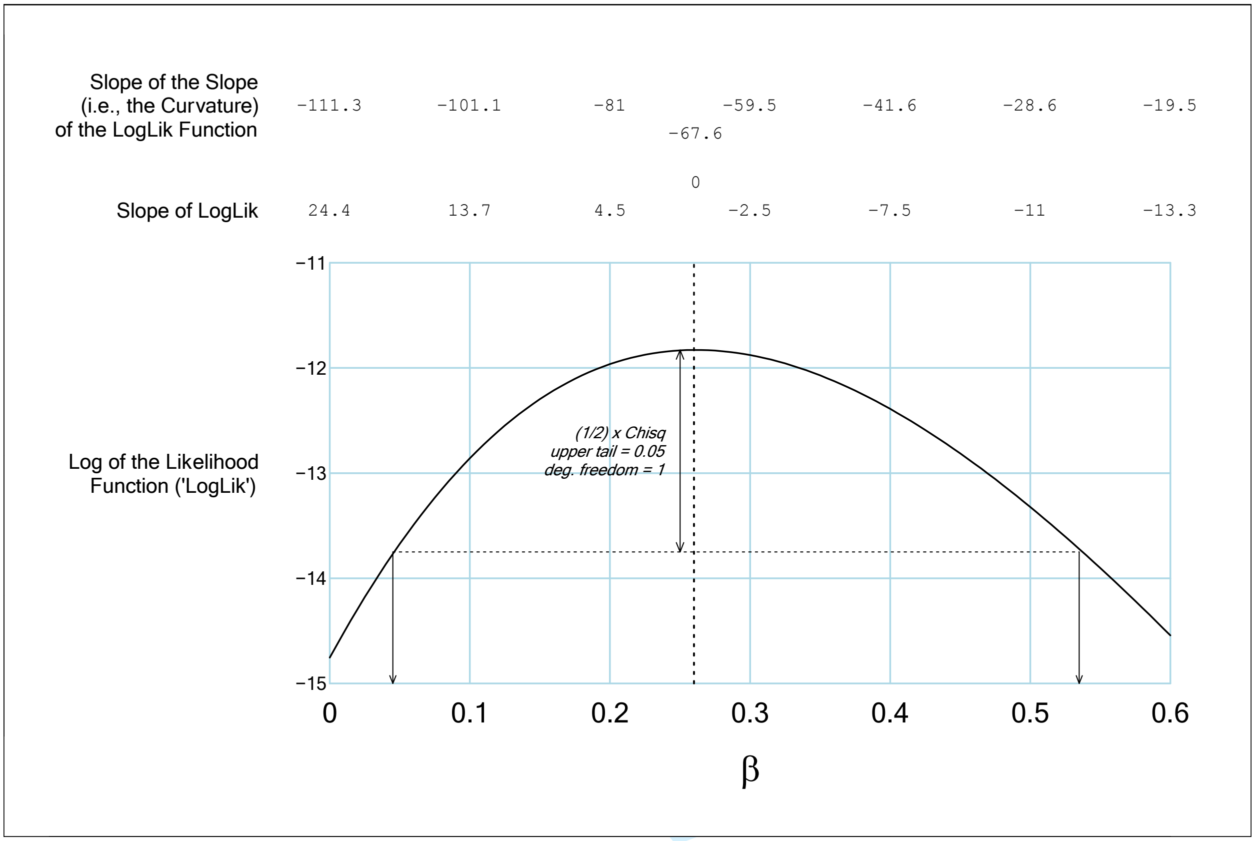


Figure A1:Log-likelihood function for the parameter β of the exposure-response model, based on the data from the 10 matched sets in figure 1, together with its first and second derivatives computed at selected parameter values. The log-likelihood function reaches its maximum at β = 0.261, where its first derivative equals 0. The quantity 67.6 measures how curved the curve is at this ML value, and the square root of its reciprocal provides the Standard Error of the fitted β. The SE can then be used to form a Gaussian-based CI, or one can use the Likelihood ratio and the Chi-Square distribution to find the range of parameter values compatible with the data (limits are marked by the 2 arrows at 0.05 and 0.53).

Before we introduce the formula-based approach that reveals where the precision (SE = 0.12) of the point estimate (0.261) comes from, we first use 'brute force' to numerically compute the standard error directly from the generic log-likelihood form. In other words, we rely *solely* on the log-likelihood *function* ('LogLik') plotted in Figure A1. The ML estimate is the parameter value at which the first derivative (slope) of the log-likelihood function crosses from positive (at the left of the maximum) to negative (at the right), namely 0.261. Its variance is the reciprocal (inverse) of the (negative of the) second derivative of log-likelihood function evaluated at this same parameter value. This makes intuitive sense: the more concentrated (the sharper, or more curved) the curve is at its maximum, the narrower is the range of parameter values supported by the data. Moreover, as we go from left to right along the β scale, the log-likelihood curve goes from low to high to low, so its *slope* (the first derivative) goes from positive to negative, and so its *curvature* (the second derivative) is negative*. The more negative the curvature is, the tighter the log-likelihood and the more precise is the point estimate.*

One can check manually/visually that the first derivative is 4.5 at $\beta = 0.2$ and -2.5 at $\beta = 0.3$. Thus, the second derivative at $\beta = 0.25$ is approximately (-2.5 – 4.5)/0.1 or -70, and so its value of -67.6 at the ML value of $\beta = 0.261$ makes sense. R.A. Fisher, who developed the ML theory in the 1920s, called the -(-67.6) = 67.6 the '*Information'* (*I*) *in the data concerning β,* and showed that *its reciprocal (i.e., 1/I = 1/67.6) can be taken as the variance of the β estimate*, so that the Standard Error, the square root of the variance, is

$$\mathrm{SE}[\,\hat{\beta}\,] = (1/\mathrm{Information})^{1/2}\ ,$$

or, in this example,

$$SE[\,\hat{\beta}\,] = (1\ /\ 67.6)^{\,1/2}\ = 0.1216.$$

One can verify that this agrees with the output from the **`clogit`** function in R or Stata or the **`phreg`** (with the **`strata`** statement) procedure in SAS.

Even though most textbooks begin their teaching of Maximum likelihood by defining the Likelihood as a *product* of probabilities, Fisher always began directly with the *log*-likelihood, so that it can be immediately written as a *sum* of the individual *log*-likelihood contributions, one from each ‘datapoint’. Quite apart from making the sum a more manageable number, the log-version immediately emphasizes that each datapoint (or matched set in our example) *adds* to the information about the parameter of concern, that not all datapoints contribute equally, and that we can readily quantify, in a technical sense, exactly how much ‘information’ each one adds.

As we will now demonstrate, by working with this formal measure of information, and just taking the reciprocal of the combined information at the very end, the factors that determine the variance and the SE of the fitted β become very clear. So, instead of relying on the numerical version of the second derivative of the entire log-likelihood function as we did above, we will now show the specific *closed-form formula* that measures the ‘information’ contributed by each riskset, using as an example that contributed by riskset 5. From equation (A1) above giving the formula for the first derivative for the log-likelihood contribution, one can use the rules of calculus to verify that the second derivative involves the same weights used in the weighed mean of the 4 temperatures, and that it is merely the negative of the weighed MSD of these 4 temperatures from that matched-set-specific weighed mean. The calculation of this weighted mean square is illustrated for selected matched set (5) in Figure 1 in the fulltext, where it is calculated under 3 scenarios: the null and ML values of β, and an intermediate value where β = 0.1. The 4 temperatures are 26, 24, 30 and 28, or, (measured from their minimum), +2, 0, +6 and +4. Thus, at β_{ML}= 0.261, so that exp(β_{ML}) = 1.3, the weights are 1.3^2 = 1.69; 1.3^0 =1; 1.3^6 = 4.79; and 1.3^4 = 2.84, so the weighted mean is 28.2. The weighted mean of the squared deviations of the 4 temperatures from this 28.2 is 4.0. As such, matched set (5) is the one with the second-smallest spread of temperatures, and it contributes the second smallest amount of information to the combined information of *I* = 67.6. The smallest contribution of the 10 matched sets is the 3.6 from riskset (4), where the temperature range was just 4.5 ˚C, and the largest is the 10.9 from set (2), where the range was 11.5 ˚C. This ranking is the same as when the information is calculated at $\beta_{NULL} = 0$. That the SE of the fitted slope is inversely related to the spread of the exposure variable makes explicit what researchers instinctively know: it is difficult to measure a slope (e.g., the fuel consumption of a vehicle) over a short distance.[10]

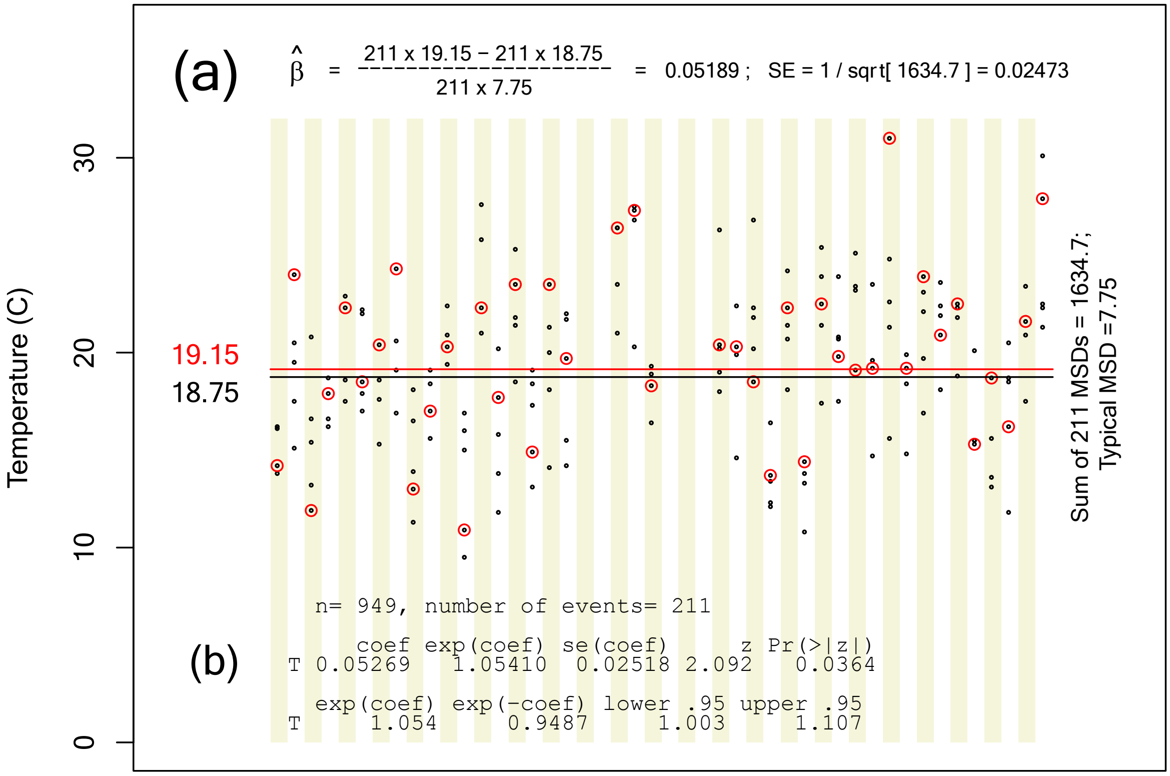
Readers may wonder why we do not refer to the weighed mean square deviation as a '*variance*'. Technically it is, but since most readers associate a variance with a divisor that is one less than the number of objects, we prefer to use the more expressive term mean square deviation. In his seminal article, Cox[11] refers to it as a "variance over the finite population of *T*'s using an 'exponentially weighed' form of sampling." This fits with the principle that in a regression model, that x's are not treated as realizations of a random variable whose variance is to be *estimated*;[9-10] the regressors are considered fixed, as if they had been decided by the investigator.

Fisher made a distinction between the *expected* information concerning β calculated using *pre-study projections* and the *observed* information calculated *post study* using the observed data. The latter is used to calculate the Standard Error for the β estimate, namely

$$SE[\hat{\beta}] = (1 / I)^{1/2} = (1 / [6.3 + 10.9 + \ldots 4.0 + \ldots + 6.8 + 8.6])^{1/2} = (1 / 67.6)^{1/2} = 0.1216.$$

Smaller signal, more matched sets

To illustrate this, we extended our case series to the 211 events that occurred in the same Canadian province during the full 30-year period for which events were documented. To more easily distinguish the matched sets, the 4/5 datapoints shown in each column of Figure A2 are the temperatures on the same day-of-week in the same-month for *every fifth one* of these 211 matched sets. To dilute the relationship, we used temperatures from another Canadian province, and used mean temperature rather than maximum temperature.



```
  n= 949, number of events= 211

      coef exp(coef) se(coef)      z Pr(>|z|)
T 0.05269   1.05410  0.02518 2.092   0.0364

  exp(coef) exp(-coef) lower .95 upper .95
T     1.054     0.9487     1.003     1.107
```

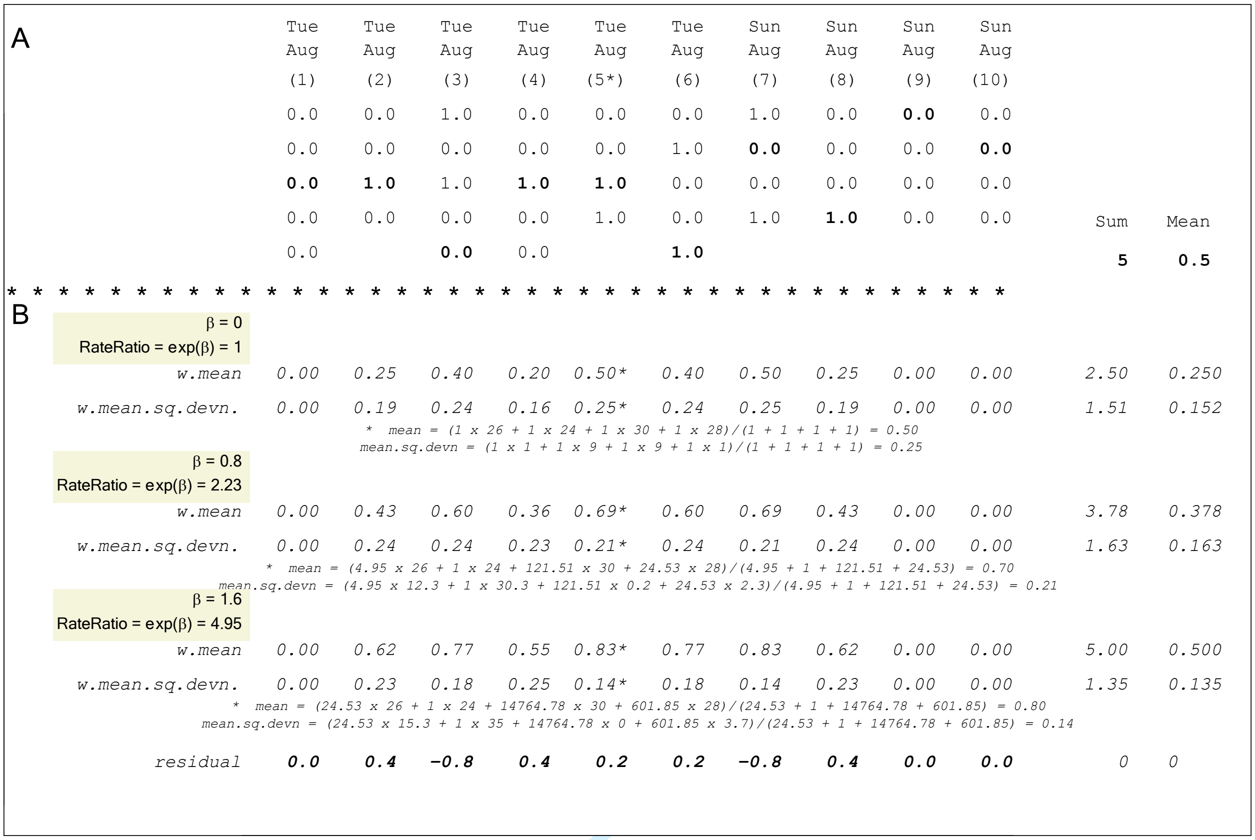
***Figure A2**: The black dots are the temperatures (T's) for every fifth one of the 211 matched sets (see text). The temperature on the day of the event is indicated by a red circle. The 19.15 on the left hand side is the mean of the temperatures for the 211 event days, while the 18.75 is the mean of the 211 matched-set means. The typical MSD is 7.75 (right hand side), and the sum of the 211 MSDs is 1634.7. The first approximation to the parameter of interest, along with its standard error (SE), is shown is shown in (a), while the ML parameter estimate and its SE (fitted via conditional logistic regression,* `clogit` *in* `R`*) are shown in (b).*

The fitted β is now 0.053, but because the SE is 0.025, the z statistic is just over 2, and very similar to the z statistic of 0.261/0.122 in our earlier example with just 10 events but a stronger signal. The greater precision is largely because of the larger number of events (211). Since the fitted β is much closer to zero, the typical MSD at the ML value (7.47) is very close to the 7.75 calculated at β = 0. Thus, the SE of 1/sqrt(7.47 × 211) = 0.0251 is only very slightly larger than the SE of 1/sqrt(1634.7) = 0.0247 calculated at the null.

Those who prefer to stay close to their data can avoid the conditional logistic regression software altogether when β is expected to be very close to 0. The first iteration of the ML procedure has the simple form shown in expression (a) in Figure A2, and yields a very good approximation to the deluxe final ML version. This very good closed form approximation in the case of weak relationships is not well known, although it was mentioned in the report[12] of a well-chronicled study of the health effects of environmental contamination.[13-14] That 1986 study had the same matched-set structure as the illustrations used here.

C A UNIFIED APPROACH TO ALL-OR-NONE AND QUNATITATIVE EXPOSURES

In our calculations thus far, there was nothing special about the fact that T is recorded on an interval scale. *Had the exposure been recorded on an all-or-none (2-point, binary) scale, the approach would have been exactly the same*: the only change would be the focus on a single Rate Ratio = exp[β] contrasting the rates in the presence and absence of the factor of interest and the 0/1 exposure scale in which the (weighted) means and squared deviations are measured. To make these ideas more concrete, Figure A2 revisits the 10 events in Figure 1, but shows a binary exposure (to stay with the same illustrative example, we merely dichotomized the temperature scale.)



A

	Tue Aug (1)	Tue Aug (2)	Tue Aug (3)	Tue Aug (4)	Tue Aug (5*)	Tue Aug (6)	Sun Aug (7)	Sun Aug (8)	Sun Aug (9)	Sun Aug (10)	Sum	Mean
	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	**0.0**	0.0		
	0.0	0.0	0.0	0.0	0.0	1.0	**0.0**	0.0	0.0	**0.0**		
	0.0	**1.0**	1.0	**1.0**	**1.0**	0.0	0.0	0.0	0.0	0.0		
	0.0	0.0	0.0	0.0	1.0	0.0	1.0	**1.0**	0.0	0.0		
	0.0		**0.0**	0.0		**1.0**					**5**	**0.5**

* *

B

	(1)	(2)	(3)	(4)	(5*)	(6)	(7)	(8)	(9)	(10)	Sum	Mean
β = 0 RateRatio = exp(β) = 1												
w.mean	*0.00*	*0.25*	*0.40*	*0.20*	*0.50**	*0.40*	*0.50*	*0.25*	*0.00*	*0.00*	*2.50*	*0.250*
w.mean.sq.devn.	*0.00*	*0.19*	*0.24*	*0.16*	*0.25**	*0.24*	*0.25*	*0.19*	*0.00*	*0.00*	*1.51*	*0.152*

** mean = (1 x 26 + 1 x 24 + 1 x 30 + 1 x 28)/(1 + 1 + 1 + 1) = 0.50*
mean.sq.devn = (1 x 1 + 1 x 9 + 1 x 9 + 1 x 1)/(1 + 1 + 1 + 1) = 0.25

	(1)	(2)	(3)	(4)	(5*)	(6)	(7)	(8)	(9)	(10)	Sum	Mean
β = 0.8 RateRatio = exp(β) = 2.23												
w.mean	*0.00*	*0.43*	*0.60*	*0.36*	*0.69**	*0.60*	*0.69*	*0.43*	*0.00*	*0.00*	*3.78*	*0.378*
w.mean.sq.devn.	*0.00*	*0.24*	*0.24*	*0.23*	*0.21**	*0.24*	*0.21*	*0.24*	*0.00*	*0.00*	*1.63*	*0.163*

** mean = (4.95 x 26 + 1 x 24 + 121.51 x 30 + 24.53 x 28)/(4.95 + 1 + 121.51 + 24.53) = 0.70*
mean.sq.devn = (4.95 x 12.3 + 1 x 30.3 + 121.51 x 0.2 + 24.53 x 2.3)/(4.95 + 1 + 121.51 + 24.53) = 0.21

	(1)	(2)	(3)	(4)	(5*)	(6)	(7)	(8)	(9)	(10)	Sum	Mean
β = 1.6 RateRatio = exp(β) = 4.95												
w.mean	*0.00*	*0.62*	*0.77*	*0.55*	*0.83**	*0.77*	*0.83*	*0.62*	*0.00*	*0.00*	*5.00*	*0.500*
w.mean.sq.devn.	*0.00*	*0.23*	*0.18*	*0.25*	*0.14**	*0.18*	*0.14*	*0.23*	*0.00*	*0.00*	*1.35*	*0.135*

** mean = (24.53 x 26 + 1 x 24 + 14764.78 x 30 + 601.85 x 28)/(24.53 + 1 + 14764.78 + 601.85) = 0.80*
mean.sq.devn = (24.53 x 15.3 + 1 x 35 + 14764.78 x 0 + 601.85 x 3.7)/(24.53 + 1 + 14764.78 + 601.85) = 0.14

	(1)	(2)	(3)	(4)	(5*)	(6)	(7)	(8)	(9)	(10)	Sum	Mean
residual	***0.0***	***0.4***	***−0.8***	***0.4***	***0.2***	***0.2***	***−0.8***	***0.4***	***0.0***	***0.0***	*0*	*0*

Figure A3: *A: Exposures, recorded on a 0/1 scale, for the day-of-week-that-month 'column' containing each of the 10 events shown in Figure 1. The exposure on the day of the event is indicated in bold. The 2 columns in which there is no variation in exposure are non-contributory. In the remaining 8, the exposure factor was present on 5 of the days when the event occurred, and was absent on 3.*

B: The calculations used in the pursuit of the Maximum Likelihood (ML) estimate of β, exactly as in Figure 2, beginning with β = 0, and ending with β = β_{ML} = 1.6. The SE for the fitted β is 1/sqrt[1.35] =0.86.

In the first two rows of Part B, readers can note one major simplification: in columns (3), (4) and (5), where the (unweighted) means of the 0s and 1s are 0.4, 0.2, and 0.5, the respective mean square deviations are the '*Bernoulli'* variances, 0.4 × 0.6 = 0.24, 0.2 × 0.8 = 0.16 and 0.5 × 0.5 = 0.25. However, they are not necessarily smaller when the weights are calculated at non-null values of β. The maximum information that any matched can provide is 0.5 × 0.5 = 0.25;

this occurs when the exposure factor is equally likely to be present/absent, and the information is calculated at the null. Further away from these situations, the contribution per matched set can be less. Thus, in equation [A] the divisors of the 1.96 and 0.84 will be much smaller than they would with a quantitative T scale (of course, in the case of a truly binary exposure, the Δ of concern would likely be larger than the 0.1 employed there).

Using the Bernoulli (and weighted Bernoulli) variances in formula A, one arrives at the same sample size suggestions as those given by the specialized packages or tables. As an example, suppose we wished to have 80% power against an alternative of β = 0.8. Again. we could use the data in Figure 2 as pilot data, and calculate that the typical information per matched set is 0.15 under the null and 0.16 under the alternative. Thus, the suggested number of events, n, is ([1.96 / sqrt(0.15) + 0.84 / sqrt(0.16)] / 0.8)^2 = 80. Using an exposure prevalence of 0.25, a Rate Ratio of exp(0.8) = 2.25, interpolation between rows 3 and 23, and columns 2 and 3, in Table 7.9 of Breslow and Day (1987), and scaling up slightly to have the average of 4.4 (rather than 1 + 4 = 5) observations per matched set, yields an n of 82.

Figure A2 shows why smaller matched sets involving a binary exposure are more likely to lack exposure variation and thus to be uninformative. A wider range of 'days' may reduce the effects of autocorrelation and increase the variation but may involve a bias/precision tradeoff.

In Table 2 of Lagakos et al.[12] the17 risksets ranged in size from 84 to 290. The proportions exposed varied from 0.18 to 0.40, so all risksets were informative. Interestingly, the authors approximated the ML parameter estimate using the closed form shown in Figure A2 (a) and intimated that the resulting β value of 1.11 may be an underestimate; in fact, the ML estimate is 0.99., and the SE at this value is smaller than it was at the null. Thus, when large effects of a binary exposure are anticipated, calculations at the null and at the alternative can be helpful to see how much information each matched set may contribute.

References

1 Clayton D, Hills M. Statistical models in epidemiology. Oxford University Press. New York. 1993.

2 Armstrong BG, Gasparrini A, Tobias A. Conditional Poisson models: a flexible alternative to conditional logistic case cross-over analysis. *BMC Medical Research Methodology* 2014, 14:122 http://www.biomedcentral.com/1471-2288/14/122

3 Cox DR. Regression models and life-tables. *J Royal Statistical Society, Series B*, 1972;34:187-220.

4 Pardoe, I. Just how predictable are the Oscars? *Chance* 2005;18:32–39.

5 Pardoe I. Predicting Oscar Winners. *Significance* 2007;4:168-173.

6 Pardoe I, Simonton DK. Applying discrete choice models to predict Academy Award winners. *J. R. Statist. Soc. A.* 2008;171:375–394.

7 McFadden, D. Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Frontiers in Econometrics*, 1974:105–142. New York: Academic Press.

8 McFadden D. Nobel Lecture. 2000 https://www.nobelprize.org/uploads/2018/06/mcfadden-lecture.pdf

9 Hanley JA and Moodie EEM. Sample Size, Precision and Power Calculations: A Unified Approach . J Biomet Biostat 2011; 2-5. http://www.medicine.mcgill.ca/epidemiology/hanley/Reprints/UniversalSampleSize.pdf

10 Hanley JA. Simple and multiple linear regression: sample size considerations. *Journal of Clinical Epidemiology* 79 (2016;79:112-119.

11 Cox DR. Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B* 1972; 34:187-220.

12 Lagakos SW, Wessen BJ, Zelen M. An Analysis of Contaminated Well Water and Health Effects in Woburn, Massachusetts. *Journal of the American Statistical Association* 1986;395:583-596.

13 Harr J. A Civil Action, Random House, New York. 1995.

14 Zaillian S. (Director). (1998). *A Civil Action* [Film]. https://www.imdb.com/title/tt0120633/

IJE Submission - IJE-2022-03-0366

International Journal of Epidemiology <onbehalfof@manuscriptcentral.com>
Fri 2022-06-10 4:47 AM

To: James Hanley, Dr. <james.hanley@mcgill.ca>

IJE-2022-03-0366
Planning and understanding sample sizes for case-crossover studies of environmental exposures

10-Jun-2022

Dear Prof. Hanley,

We have now had a chance to review the paper you submitted to the International Journal of Epidemiology. The paper has been refereed by at least one external reviewer and has also been read by an Editor or Associate Editor of the Journal.

I am afraid that we found that your submission was not suitable for publication in the International Journal of Epidemiology. This decision was made both in response to referee comments and also on the grounds of suitability with respect to the journal and priority in relation to the other submissions we have received. The decision is thus not solely influenced by the particular comments of referee(s).

We enclose a set of referee comments which we hope will be of use when revising your paper for submission elsewhere.

Thank you for your interest in the International Journal of Epidemiology.

Yours sincerely,

Stephen Leeder
Editor-in-Chief

Comments from Editor:
Four biostatistical reviewers have now gone over this manuscript and while they agree on the importance of the topic, the paper is unfortunately not suitable for Education Corner. Most importantly for the pedagogical goals of Education Corner: the paper covers only one aspect/method without setting in the context of other approaches or other decision making factors, such as when sample size calculations are warranted (e.g. costly primary data collection) vs. are not warranted (e.g. secondary data analyses of administrative data) for observational environmental epidemiology.

The reviewers appreciated the technical aspects of the proposed method and the case studies, but noted some inconsistencies in presentation where, e.g. some equations were presented without sufficient elaboration and other important details were relegated to appendices.

The reviewers provided substantial, detailed comments that we hope will be helpful to you in submitting this work elsewhere.

Comments from Referees:

Referee: 1

Comments to the Author
This manuscript submitted for the IJE Education Corner section aims at planning sample sizes for time stratified case-crossover studies using categorical or continuous exposures (environmental or not). Case crossover designs (with a time-stratified setting or more modern approaches) are used extensively in environmental epidemiology and other fields focusing on short-term exposures. There are many papers (some of them cited in this paper) discussing the sample size considerations for such design.

While I found the discussion of the ML estimation applied to this issue very interesting, I think this paper is out of the scope of IJE and especially as an educational piece. It reads more like a statistical note on a very specific aspect of such study design and does not propose a comprehensive roadmap for sample calculation for case crossover designs. Discussions about the pertinence of such a priori power calculations when using observational data is lacking as I wonder in which case such calculations can help an investigator considering such method and potential implications (e.g. collecting more data, using alternative methods, abandoning the study…).

In summary, this manuscript addresses a specific a priori power question and I don’t think this paper could be helpful for readers that are not already familiar with the case crossover estimation details. Furthermore, there are multiple considerations that are not discussed, including complex lagged and time-varying confounding structures. The introduction and the general context sections seem to be written in haste and does not provide enough details of more recent developments in this literature.

Referee: 2

Comments to the Author
The manuscript describes power and sample size computation for studies applying the case-crossover design, together with algebraic definitions. The topic is of interest, with interesting discussions on several aspects and an illustration of real-data applications. However, in the current form, the manuscript is very hard to follow, with superficial information provided in the main text, a lack of description of the basic design settings, and confusion due to the use of different data examples. Detailed comments are provided below.

1. I found the structure chosen for this contribution very confusing. First, the authors provide limited information in the manuscript about the design per se (see Comments 7-9 below) and the statistical quantities (see Comments 11-12), with most of the latter confined to the appendix. Second, the authors use two different data examples, one very basic for the main paper and one more structured and coming from a real dataset in the appendix. Third, the appendix includes a

long description of the underpinning likelihood theory, which, while interesting, is not essential. These issues result in a very superficial description being offered just by reading the article, which is not well complemented by the very complex theoretical overview provided in the appendix. I suggest the authors consider revising the article using a different structure and content.

2. One very simple way of doing it would be to use a single real-data example (I would recommend the tornado data), and illustrate all the examples and steps using it. For instance, the results of the equations for the minimum sample size and power can be computed for this (relatively) small dataset, as well as the MSD/MSW and the likelihood contribution for each risk set.

3. Similarly, the algebraic definitions of the MSD/MSW should be provided in the main text, as they are central to the power calculations.

4. I am not sure all the information provided in the appendix is relevant. For instance, most of the likelihood theory (part B) is interesting, but very general and not specific to this context. If the authors choose to keep it, I would move it to the very end of the appendix and focus first on more specific aspects related to the case-crossover design.

5. The provision of a code, for instance using the R software, would enormously facilitate the application of these methodologies by the users. I suggest providing a script and real-data example, possibly replicating the results described in the manuscript.

6. The authors chose to focus their contribution on the application of the case-crossover design in environmental research. However, exactly the same models and power calculations for this design can be used in other epidemiological areas. I wonder if the authors can make the presentation more general, and use the application for studying risks associated with environmental exposures only as an example. This would make the contribution relevant to a broader audience.

7. The authors need to contextualise the use of power calculations in this setting. In the majority of the cases, the data collection and analysis of a case-crossover study do not require the drafting of a pre-specified protocol and are easy to perform. This does not mean that power calculation is not required, but the authors should clarify at which step of a project a researcher can find it useful.

8. The authors provide very little detail on the structure of the case-crossover design, and given the type of contribution (Education Corner), it cannot be assumed that all the readers are familiar with it. In particular, several control sampling schemes exist, and the authors only describe the most commonly applied in environmental epidemiology, i.e. the time-stratified with month/weekday strata. The authors should provide a more general presentation, describing the general structure of the design and motivating the decision about presenting this specific scheme.

9. There are very few references in the article. The authors can include up to 30 of them, and I strongly suggest adding more.

10. For the description of the method, if the authors choose to follow my suggestion in Comment 1, I would refer to an 'exposure of interest x' rather than temperature or any specific factor. This can make the illustration more general. The authors can then add a specific example using real data.

11. The description of the statistical model and the quantities of interest are confused. For example, in Equation 1 the outcome lambda does not represent a rate (e.g., number of events per day), but an individual hazard, as in Cox proportional hazard model. The case-crossover design works with individual data. It is true that with aggregated data its likelihood can be replicated using conditional Poisson models (see reference 2 in the appendix), which work with rates. However, in this case, the analysis is not performed using a logistic regression, which only accepts a binary response and not a count. I suggest the authors describe the individual-level setting and then mention the case of aggregated data as a specific extension.

12. The authors should make clearer to the reader what are the theoretical foundation of the separation in weaker and stronger-relationship scenarios. My understanding is that it is related to the possibility of making an approximation in the weaker-scenario setting, but is it unclear how the threshold of an RR. I could not find a clear explanation neither in the main manuscript or in the appendix.

13. There is another complication that is not addressed in this contribution. The typical case-crossover setting can be cast as a nested case-control in which a selection of potential control-days are taken in each risk set. For instance, in the time-stratified sampling scheme, the same weekdays in the months are selected, although it is possible to select all the other days in the month and to control by day of the week directly (a sort of full-stratum scheme). In this case, it can be expected that the latter method provides more power, as more controls are selected for each case. The authors can refer to power calculation in nested case-control studies for references. I assume that this is not clear in this contribution as the authors keep referring to an aggregated data structure, in which days and not individual events are the unit of analysis. In this case, days are repeatedly taken both as cases and controls, and therefore the problem above does not apply. However, for individual-level analyses, I assume that another parameter that influence power is the number of control taken in each risk set. The authors should at least mention this issue.

14. It would be good that the authors clearly state somewhere in the article that the estimators from logistic regression in case-crossover design define a log risk ratio, not a log odds ratio. This is implied in some parts when describing the method, but never clearly stated. I think this is not clear to the majority of users of the design.

15. I strongly suggest defining the multiplier for the confidence level explicitly, instead than approximating it to 2 for the 95% option. Similarly, the concepts of type-I and type-II errors should be described, and the related quantities (e.g., alpha and 1-alpha) defined. For instance, it is not clear to the reader what Z_beta=0.94 really is on page 9.

16. The 'margin of error' seems the width of the confidence interval. If so, I would refer to it explicitly.

17. Add numbers for all the equations.

18. The title is not appropriate. I would refer to ‘power and sample size’.

Referee: 3

Comments to the Author
Summary points
The paper discusses important considerations for adequate power in time-stratified case-crossover studies, which are commonly applied to the study of acute outcomes and exposures such as heat and air pollution. The paper highlights key statistical considerations, and walks readers through a worked example for a sample size calculation. This paper is a welcome addition to the case-crossover literature, which lacks easy to implement sample size calculations. This work will help investigators ensure time-stratified case-crossover studies are adequately powered to detect an effect. However, additional work is needed to increase clarity and improve organization to ensure understanding of the key points.

General comments
1. Overall, this paper first comes across as easy to follow. However, while overall the language is simple, it is inconsistently simple and, in some cases, this is problematic because of its lack of precision and clarity. Further, not all terms are clearly defined.

2. The paper does not use much of the standard terminology in the case-crossover literature, i.e., it omits words such as “index” and “referent” day (or time), “referent window” and “stratum”. Thus, it makes it more difficult for readers of this paper to connect its content with the methodological papers that have already been published on the topic. This is a disservice to readers of this paper.

3. Please use standard capitalization of terminology. (Examples: don’t capitalize maximum likelihood, mean square deviation, least squares.)

4. The organization of the paper should be revisited. In particular, it appears that the material in Appendix A was moved to the appendix late in the manuscript drafting stage. There is reference to the example in this Appendix in the main text, but it is not appropriately referenced so the reader doesn’t know what the authors are referring to. Further, there is considerable text in the appendix covering binary exposures, but no reference to this in the main paper.

5. The source of every sample size formula should be referenced. One source is on page 9, although the formula in the reference is sufficiently different from the one given on page 9 that further explanation by the authors is needed in order for readers to make sense of these.

6. Basic definitions should be given. For instance, the variance of interest for all the sample size examples is 1/(MSD*n) where n is the number of events. This is never defined and leaves the reader to just trust the authors claims about various formulas without allowing deeper understanding.

7. The authors should reference other papers that discuss case-crossover study sample size calculations and state how their contribution is distinct. For instance: Dharmarajan, S, Lee, J-Y, Izem, R. Sample size estimation for case-crossover studies. Statistics in Medicine. 2019; 38: 956– 968. DOI: 10.1002/sim.8030

Specific comments
Abstract
1. Abstract is a helpful and adequately detailed summary of the paper.
2. The key messages also nicely convey the key takeaways, with one exception (see below)
3. P 1 Line 37: Add "of time-varying" before "environmental"
4. P 2 Line 16: While the same considerations may apply to non-environmental exposures, can the authors think of an example where the time-stratified design would apply? If not, then I think this sentence should be dropped. If so, it would be good to mention this in the paper (though it is too detailed for the abstract). Also drop this from the key points, unless the authors can support this statement better in the paper.

Introduction
1. Consider adding a sentence or two about the benefits and/or disadvantages of the case-crossover study design compared to alternatives, such as Poisson or quasi-Poisson time series analyses. While readers will likely be familiar with the design, it may provide helpful context.
2. P 4, L 22-28: The description here seems loose and also assumes that the appropriate time unit is day. It will also be useful to incorporate the idea of a stratum or referent window here when talking about "days with similar characteristics". Part of the idea is that strata can (and should) be defined in advance without looking at the data.
3. P 4 L 45: Suggest replacing "fit" with the more precise term "estimate"
4. P. 5, L 21: change so the word "typical" is not repeated

Preliminaries
1. This section is for the most part clearly communicated and easy to follow
2. P 6 L 34: It will probably help some readers to express this statement mathematically, e.g. exp[\beta(T_max – T_min)] < 1.1, where T_max and T_min are the maximum and minimum temperature within a referent window
3. P 7 L 31: Suggest adding "in many environmental exposure settings" after "common"

Scenario sections
1. These sections are easy to follow. Providing 2 scenarios (one more common, but with a weaker relationship, and with a stronger relationship) is helpful for solidifying comprehension
2. P7 L 43: Is this section, "Number of cases to ensure a desired precision", useful to include? When would investigators want to design a study based merely on precision and not on power?
3. P 7 L 51: Since the authors are so careful to define many basic terms, I suggest defining margin of error also. Further, it is difficult for readers to connect this section with the next one without this ME definition explicitly stated since the formula given here doesn't include the effect size of interest.
4. P 8 L 7-11: It is not clear what is the basis of the statement and formula. Why not explain this? (Assuming the section is kept). Further, the authors only implicitly equate "its standard error (SE)" on line 4 with "SE_desired" in the formula. This should be explicit. It would also be helpful for there to be a definition of the variance term of interest, which is 1/MSD. The

explanation that starts on line 15 is helpful; an additional connection is what is needed.
5. P 8 L 51: Don't the authors mean cut the margin of error in half? Spelling out the words four and half will increase clarity.
6. P 9: The framing of the section "number of cases to ensure a specified power" is confusing. Ensuring adequate power is asking a different question than merely focusing on targeting precision. The "added insistence" language is trivializing this feature.
7. P 9 L 8-11: The justification for this sentence is absent given the precision section didn't focus on effect size whereas this is an inherent feature of the sample size formula in this section.
8. P 9 L 21: Note that the beta in Z_beta is in reference to something different than the beta used elsewhere in the paper. Some readers will be confused if this is not explained.
9. P. 10, L 15: MSW is referenced. Do you mean MSD here or MSW? Please write this out the first time it is referenced and the term is not defined.
10. P 10 l 38: The reference to matched set 5 refers to the example in the appendix, which hasn't been introduced yet, so the reader doesn't know what "matched set 5" is.
11. P 11, L 23: Several of the terms used in this equation are not defined. Further, the authors should comment on why the SEs might be different for the null and alternative hypotheses and when in practice this should be considered.
12. In the stronger relationship scenario the authors use SE = 1/sqrt(MSW*n), where it appears that MSW (which isn't defined in the paper) = MSD. This usage seems to be consistent throughout the paper, but is never explicitly defined.

Discussion
1. P 12, L 50+: The authors are making an important point, but the example risks consumers thinking that the MSD for all strata will be the same as it is in one stratum. It is important to emphasize that the within-matched-set variation is an average across strata. As the authors imply, this is less that the overall variation, at least when the mean exposure varies from stratum to stratum.
2. P. 13, L 21: Provide guidance or an example of a strong exposure-response relationship where this method may not suffice
3. P. 13, L 36: Clarify what 'them' refers to
4. P. 13: first and second full paragraphs on this page do not flow well together. Consider revising
5. P 14 L 4+. This sentence doesn't make sense
6. P 14 L 11: This is a throw-away comment that should be developed more or dropped. Please present an example of a case-crossover study of non-environmental exposures that would use the time-stratified design or drop this.
7. Here or elsewhere, consider adding discussion of how this method translates to case-crossover studies with lagged exposures

Appendix/supplement
1. Add references to the supplement sections in the main text so as to make it easier for readers find additional detail on a topic as needed
2. Throughout the supplement, considering streamlining to include only information relevant to the goal of this paper
3. Please drop the "/" usage, e.g. mean/sum, fitted/weighted, 4/5. It is really confusing, unnecessarily so.

4. P 15 L 25+: This phrasing doesn’t make sense. Tornadoes are the outcome and the scientific question of interest is whether the rate of tornadoes increases with temperature.
5. Figure 1 should be more clearly documented. For instance,
a. How are the columns in A related to T?
b. How are the strata defined?
c. What is the rationale for the rescaled weights?
d. Add text to indicate part A shows the data.
e. Use notation that incorporates indexing of the stratum and the observations within stratum, e.g. Ti,j for stratum i and observation j.
f. Make it clear that the events are indicated by boldface type.
g. Explain why the weights are the same for beta=1
h. Where are the fitted/weighted means defined?
i. Why is there new terminology in this figure (e.g., w.mean.sq.devn) that isn’t used in this paper?
6. p 17 l 12: Is this (“section B”) related to Figure 1 or the next part of the Appendix?
7. P 17 L 22: Define ‘fitted’ T’s here or earlier
8. P. 17, L 25. This is the supplement. Drop the phrase “As we explain in the supplement”. Also MSD isn’t connected readily to what is shown in Figure 1.
9. P 17 L 39+. Figure 1 presents results for 3 betas. The reference to beta in this sentence is unnecessarily unclear.
10. P 18 L 18+: This sentence is confusing and seems counter-intuitive. As a beta gets larger the required sample size to detect it ought to be smaller, even if the SE gets smaller. Please address.
11. Section B:
a. Why start with a definition of multinomial probabilities?
b. P 19 L 34: The notation is odd and very confusing. The numerator is a vector (though nontraditionally expressed with curly brackets) and the denominator is a sum. Please define the notation clearly.
c. P 19 L 39+. This statement may be correct, but each example assumes a model and that is lost. It is bound to confuse readers
d. P 20 first sentence: Why talk about a sine curve and specifically what discrepancies are being referred to?
e. P 20: Why is maximum likelihood very different? Aren’t the “discrepancies” connected to the “balancing equation”?
f. P 21 L 44: I think it would be clearer to state that the ratio is a form of 1 and doesn’t change the equation. Or show it. The wording can be misunderstood.
g. P 25 L 35+: I think the slopes should be shown on Figure A1 to improve clarity
h. P 26 L 12+. Why not show code for this? If this is supposed to be a helpful tutorial paper, it seems like the code should be included in the supplement, at least for the tornado example dataset.
i. P 27 L 24: What full text?
j. P 27 l 27: Why is values plural? Why not say “ML value of beta = 0.261” here?
k. P 29 L 14: Why not say that each column shows the 4 or 5 data points for a single stratum in the dataset where only one in every five strata is depicted?
l. P 29 L 21: Please start the section on this page with this concept.
m. P 31 l 10: Largely or entirely??
12. P. 32, L 24. This should be Figure A3 (rather than Figure A2)

13. P. 33, L 55. Missing word. Should be “any matched set can…”
14. The binary exposure example is helpful. In this setting, was an adequate range of proportion of risk sets exposed, and with exposure variation? Further commentary would be helpful, specifically stating that the strata with no exposure variation don’t contribute any information to the analysis. (This is implied at the top of p. 35, but could be developed more.)
15. P. 35, L4. Referenced figure should be Figure A3.
16. P 35, L 7: This statement isn’t well supported and is confusing. What is meant by “days” here?
17. Consider including example R code in the appendix to facilitate easier implementation of these methods (as mentioned above)

Environmental Epidemiology

Planning and understanding sample sizes for case-crossover studies of environmental exposures

--Manuscript Draft--

Manuscript Number:	
Full Title:	Planning and understanding sample sizes for case-crossover studies of environmental exposures
Article Type:	Original Research Article
Keywords:	Parameter estimation; Conditional Logistic Regression; Exposure Variation; risksets; matched sets
Corresponding Author:	James Hanley, Ph.D. McGill University Montreal, QC CANADA
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	McGill University
Corresponding Author's Secondary Institution:	
First Author:	James Hanley, Ph.D.
First Author Secondary Information:	
Order of Authors:	James Hanley, Ph.D.
Order of Authors Secondary Information:	
Manuscript Region of Origin:	CANADA
Abstract:	Background : Time-stratified case-crossover studies are increasingly used to quantify the relationship between the rates of acute health events and levels of environmental exposures such as heat or air pollution. When exposures are to be measured on a continuous scale, few sample-size planning tools are available to anticipate the statistical precision of the resulting effect estimate, or to appreciate the study design aspects that influence statistical power. Methods: We provide formulae that can be used to plan the sizes of time-stratified case-crossover studies with exposures measured on either a categorical/continuous scale. We explain where the formulae come from using a small hand-worked example. In a supplement we illustrate the calculations involved in estimating parameters from the relevant conditional logistic regression models. The expected amount of statistical 'information' that each matched set contributes to the parameter estimate is emphasized. Results : The precision of the estimated regression coefficient in time-stratified case-crossover studies depends on both the number of cases studied and the variation in the exposure values within a typical matched set (as measured by the Mean Squared Deviation). Importantly, the within-matched-set variation in continuous exposures will often be much less than the variance of exposures observed over the duration of the study period (e.g., daily variations in outdoor temperatures during 2022 compared with variation of daily temperatures for all Fridays in July 2022). Conclusions : Investigators conducting such studies should pay close attention to the expected within-set variation in exposures to ensure that an adequate number of cases is identified.

2022.07.12

Editor: Environmental Epidemiology

Dear Editor

We are submitting a piece entitle “Planning and understanding sample sizes for case-crossover studies of environmental exposures” for consideration in Environmental Epidemiology.

My co-author is quite experienced in environmental epidemiology, but tells me that he is constantly being asked by grant review panels to justify the sizes of his studies and associated budgets, but that he is unable to cite suitable sources. For such a widely used design, this is embarrassing. It also points to a common problem in epidemiology: textbooks treat every study design as a separate ‘silo’ and don’t emphasize the statistical connections between them. Of course, commercial firms that market software for ‘sample size planning’ have a vested interest in keeping these markets separate, and maintaining ‘black boxes.’

Our piece fills that gap, and without the need for anything more than a hand calculator. In addition, in an appendix that we designed to be an online supplement, it provides the ‘why’ behind the formulae and de-mystifies the algebra. That appendix also shows that we don’t need separate considerations for binary or categorical exposures: they all have a common structure/anatomy.

We hope you agree that the teaching approach in our piece makes it a suitable candidate for Environmental Epidemiology.

James A. Hanley
and Scott Weichenthal

2022.07/12 For: Environmental Epidemiology

Type: Original Research Article

Title: Planning and understanding sample sizes for case-crossover studies of environmental exposures

Authors
James A. Hanley[1*]and Scott Weichenthal[1]
[1]Department of Epidemiology, Biostatistics, and Occupational Health,
McGill University, Montreal, Canada

***Corresponding Author.**
Department of Epidemiology, Biostatistics, and Occupational Health. McGill University, Montreal, H3A 1G1, Canada. E-mail: James.Hanley@mcgill.ca

Suggested running head: sample sizes for case-crossover studies

Conflicts of interest*:* None declared.

Sources of funding*:* This work was supported by CIHR.

Ethical Approval of Studies/Informed Consent*:* Not applicable: data used are meteorological only and are in the public domain.

Planning and understanding sample sizes for case-crossover studies of environmental exposures

James A. Hanley[1*]and Scott Weichenthal[1]
[1]Department of Epidemiology, Biostatistics, and Occupational Health,
McGill University, Montreal, Canada

*Corresponding author. Department of Epidemiology, Biostatistics, and Occupational Health. McGill University, Montreal, H3A 1G1, Canada. E-mail: James.Hanley@mcgill.ca

Abstract

Background: Time-stratified case-crossover studies are increasingly used to quantify the relationship between the rates of acute health events and levels of environmental exposures such as heat or air pollution. Especially when exposures are to be measured on a continuous scale, few sample-size planning tools are available to anticipate the statistical precision of the resulting effect estimate, or to appreciate the study design aspects that influence statistical power.

Methods: We develop and provide formulae that can be used to plan the sizes of time-stratified case-crossover studies with exposures measured on either a categorical or continuous scale. We explain where the formulae come from using a small hand-worked example. In a supplement we illustrate the Maximum Likelihood (ML) calculations involved in estimating parameters from the relevant conditional logistic regression models. The expected amount of statistical 'information' that each matched set contributes to the parameter estimate is emphasized.

Results: The precision of the estimated regression coefficient in time-stratified case-crossover studies depends on both the number of cases studied and the variation in the exposure values within a typical matched set (as measured by the Mean Squared Deviation). Importantly, the within-matched-set variation in continuous exposures will often be much less than the variance of exposures observed over the duration of the study period (e.g., daily variations in outdoor temperatures during 2022 compared with variation of daily temperatures for all Fridays in July 2022).

Conclusions: Investigators conducting such studies should pay close attention to the expected within-set variation in exposures to ensure that an adequate number of cases is identified.

Keywords:

Parameter estimation; Conditional Logistic Regression; Exposure Variation; risksets; matched sets

Word count: 2708

Key Messages What this study adds

It provides simple (but unavailable until now) formulae that can be used to plan the sizes of time-stratified case-crossover studies with exposures measured on either a categorical or continuous scale.

It emphasizes that the precision of the estimated regression coefficient in time-stratified case-crossover studies depends on both the number of cases studied and the variation in the exposure values within a typical matched set (as measured by the Mean Squared Deviation).

It urges Investigators to pay close attention to the expected within-set variation in exposures to ensure that an adequate number of cases is identified.

The basis for the formulae can be understood by working through a small hand-worked example.

Introduction

Time-stratified case-crossover studies are commonly used to estimate the acute health impacts of environmental exposures such as heat or air pollution.[1] This design begins by identifying cases (e.g., *instances* of persons admitted to hospital for a myocardial infarction). For each instance, exposure data are obtained for the day of the event (or a time-period immediately prior to it), and for other days with similar characteristics (i.e., the same day of the week, month, and year). For example, if an event occurred on a Friday in July 2022 in a certain region/area, the exposure data pertaining to this case might be assembled for all Fridays in July 2022 in this region/area. We will refer to this *set of days* (which *includes* the day the event occurred) as a *matched-set***.** The matched-set is the conceptual counterpart of the risk-set in a survival analysis, or in a standard incidence-density-based matched case-control study. Typically, conditional logistic regression models are used to fit the event rate as a function of the exposure level.

In the initial applications of the case-crossover study design, the focus was on personal triggers of events, and thus the unit of observation is one *person*. Conceptually, the parameter is a rate or hazard ratio contrasting days that persons were exposed and unexposed (or if the exposure is quantitative, between levels on different days). Any important personal circumstances of each person who suffered the event that differ between the person's exposed and unexposed days are recorded and taken into account.

However, in many of time-stratified case-crossover studies in environmental epidemiology, the unit of observation is (effectively) the entire *population* on the days in question in the geographic region where the event occurred. Over the days in question within a region from which the case arose, the population (which is not enumerated or described) is assumed to remain constant in size and composition, and all persons in it share the *same* level of exposure. Other than the age and sex of the person who suffered the event, and possibly some unchanging characteristics of the region (which can only be used to study effect modification), no other covariates are recorded. Daily (or possibly finer) variations in exposure level (or levels, if exposure has more than 1 dimension) within the case region become the exposure (or triggering) variable of interest. The formulae in the main article apply when the measured exposure is 1-dimensional; the Supplement Digital Content addresses the multi-dimensional case.

Despite the common use of the case-crossover design in environmental epidemiology, and despite the fact that it can be viewed as a simpler version of the more general case-crossover design, guidance on factors that determine sample size and statistical power in these environmental studies is not readily available. To address this gap, we provide formulae that can be used to calculate these quantities and illustrate the theory behind these equations using a simple hand-worked example. We emphasize that the amount of statistical 'information' each case contributes to the parameter estimate can be quantified by the typical Mean Square Deviation (MSD) within a typical matched set. If this MSD is expected to be small, a larger number of cases must be included. Although our focus is on continuous exposure measures typical of environmental epidemiology, the same statistical principles apply to binary or categorical exposures, and to non-environmental exposures. Further technical details and explanations are provided in the Supplemental Digital Content.

Preliminaries

The parameter of interest, i.e., the estimand

For concreteness, we begin with an example where some aspect (e.g., mean, maximum) of the daily temperature is the (locally shared) exposure of interest. Thus, without loss of generality, we simply refer to it as *'T'* rather than the statistical regressor *X*, or generic epidemiological *E*. Suppose the model we will use to relate the event rate (i.e., the expected number of events per day) to *T*, is

$$\lambda(T) = \lambda_0 \times \exp[\, \beta\, (T - T_0\ )\,]\,, \tag{1}$$

where λ_0 refers to the event rate at some reference temperature, T_0 and the shorthand 'exp' stands for 'the exponentiated value of.' Thus, $\exp[\,\beta\,(T - T_0)\,]$ denotes the ratio of the event rate at temperature *T* to the rate at the reference temperature; if *T* is measured in degrees Celsius, then β, the parameter to be fitted to (estimated from) the data, refers to the log of the ratio of the event rates at temperatures that are 1° Celsius apart. Thus, ultimately, the estimand is β or exp[β].

Since different regions have populations of different sizes and compositions, the reader may wonder why these are not included in equation (1) and why the specification of the expected number of events per day is somewhat incomplete. The answer is that while these demographic factors could be in principle ve included in a more general equation, the matching (conditioning) on region, day of week, etc, means that they to cancel out in the rate ratio (the estimand).

Moreover, since demographic information on each region is not typically recorded as part of the case-crossover study, and since regions and times of the year that do not produce events are not even considered, it is not possible to fit rate-difference models.

For reasons that will become evident later, we will divide our presentation into two scenarios, which we *arbitrarily* divide into 'weaker' and 'stronger' exposure-response relationships. By 'weaker' we mean a coefficient β such that, over the *T* range in a typical matched set, the rate at the upper end is less than 1.1 times the reference rate (of 1) at the lower end. As we will see below, the sample size calculations in the 'weak' scenario are considerably simpler.

The spread of the exposure data in a typical matched set

Suppose that for a typical matched set of 4 days in a time-stratified case-crossover study, the exposures (values of *T*) for the 4 days in the set are: 21°C, 23°C, 25°C and 27°C. Of course, temperatures will not usually be so rounded or so regular: these 4 values were selected to make for convenient calculations. The mean of these values is 24 and the <u>m</u>ean <u>s</u>quared <u>d</u>eviation (MSD) from 24 is 5 ($[(24\text{-}21)^2 + (24\text{-}23)^2 + (24\text{-}25)^2 + (24\text{-}27)^2] / 4 = (9+1+1+9)/4 = 20/4 = 5$. Note that the MSD of 5 is the same as if we had recoded the 4 *T*s as 0, 2, 4 and 6°C above the minimum in the set. It is important to note that this MSD is smaller than the typically calculated *sample variance* of the 4 values. In the case-crossover context the sum of the 4 squared deviations is divided by 4, not 3, since we are not *estimating* a population variance, but rather measuring how spread out the 4 *T*s are.

With these preliminaries, we first address the number of cases (and thus the number of matched sets) to ensure that the regression coefficient β will be estimated with a specified level of precision. Since the ‘weaker-relationship’ scenario is more common, we begin with it; as it happens, the calculations in this context are also simpler.

Weaker-relationship scenario

Number of cases to ensure a desired precision

Suppose that we set the precision with which the regression coefficient β will be estimated by specifying that its 95% margin of error (ME) will not exceed some specified amount. This implies that its Standard Error (SE) will not exceed 1/2 (technically 1/1.96) of this ME. The number (*n*) of events required to achieve this *SE* is given by the formula

$$n = \frac{1}{(\,SE_{desired})^2} \times \frac{1}{Mean\ Sq.Deviation\ of\ exposure\ values\ in\ a\ Typical\ Matched\ Set}. \qquad (2)$$

It makes sense that the MSD is in the *denominator* of the formula, just as it is in the expression for the variance of a. fitted slope in a simple regression, where the narrower/wider the spread of the x’s the more/less stable will be the fitted slope.^{2} As an example, suppose the *T*’s in a typical matched are expected to have a MSD of 5. Suppose that, relative to the reference T_0, the anticipated rate ratio at $T_0 + 1$ is 1.05, so that β = ln(1.05) = 0.049. Suppose we wish the SE for the fitted β to be no larger than 0.02 (or that the margin of error not exceed 0.04). Then, to achieve this, equation (2) indicates that we would need to study

$$n = \frac{1}{(0.02)^2} \times \frac{1}{5} = 500 \text{ events.}$$

If we we wish the SE to be no larger than 0.01 (or the margin of error to not exceed 0.02), we would need to study $n = (1/0.01)^2 / 5 = 2{,}000$ events, i.e., it takes 4 times as many cases to cut the margin of error in 2.

Number of cases to ensure a specified power

The sample size to guarantee a pre-specified power (of say 80%) is larger than when (in the absence of null hypothesis testing) precision is the only concern. The larger requirement stems from the added insistence on an 80% chance that (under the alternative) the point estimate exceeds the criterion for a 'statistically significant' result. Typically $Z_{\alpha/2} = 1.96$ for a 2-sided test with $\alpha = 0.05$, and $Z_{\beta} = 0.84$ for 80% power. Thus, if Δ is the difference between the alternative and null values of β, the required number of events n is

$$n = \frac{(1.96+0.84)^2}{\Delta^2} \times \frac{1}{Mean\ Squared\ Deviation\ of\ Exposures\ in\ Typical\ Matched\ Set} \tag{3}$$

In our example, if the alternative (to the null $\beta = 0$) is $\beta = 0.049$, and the anticipated Mean Squared Deviation of the T's in a typical matched set is 5, then equation (3) indicates

$$n = \frac{2.8^2}{0.049^2} \times \frac{1}{5} = 653 \text{ events.}$$

The 'anatomy' of this formula is similar to that of equation (5) in a not-well known but quite instructive 1985 article.[3]

Stronger-relationship scenario

Number of cases to ensure a desired precision

For a desired degree of precision, the formula has the same form as the earlier one, except that each MSW is now a *weighted* MSW, and is thus somewhat narrower that the un-weighed one.

$$n = \frac{1}{(SE_{desired})^2} \times \frac{1}{Weighted\ Mean\ Sq.Deviation\ of\ exposure\ values\ in\ a\ Typical\ Matched\ Set}. \quad (4)$$

Why it is smaller is explained in the Supplemental Digital Content. However, since the β is larger than in our initial example, the stronger signal means that the required number of events may be smaller. To make these aspects concrete, suppose that, relative to the reference T_0, the anticipated rate ratio at $T_0 + 1$ is 1.2, so that β = ln(1.2) = 0.18. Suppose we wish the SE for the fitted β to be no larger than 0.075 (or that the margin of error not exceed 0.15). To get a sense of a typical weighted MSD, consider the data in Figure 1A, and treat them as if they came from a pilot study.

A

Thu May	Sun Jun	Sat Jun	Tue Jul	Wed Jul	Mon Jul	Sun Aug	Tue Aug	Thu Sep	Fri Sep
(1)	(2)	(3)	(4)	(5*)	(6)	(7)	(8)	(9)	(10)
15.0	25.0	28.0	26.5	26.0	26.5	29.0	20.5	**24.0**	20.0
23.0	18.5	20.0	24.0	24.0	27.5	**26.0**	23.5	19.5	**26.5**
13.5	**30.0**	29.0	**28.5**	**30.0**	26.5	20.0	23.5	15.0	16.0
20.0	21.5	25.5	26.0	28.0	21.5	29.5	**29.0**	20.0	22.5
20.5		**22.5**	22.5		**32.0**				

B

β = 0 RateRatio = exp(β) = 1										
w.mean	*18.4*	*23.8*	*25.0*	*25.5*	*27.0**	*26.8*	*26.1*	*24.1*	*19.6*	*21.2*
w.mean.sq.devn.	*12.7*	*18.3*	*11.3*	*4.3*	*5.0**	*11.2*	*14.3*	*9.4*	*10.2*	*14.6*

** mean = (1 x 26 + 1 x 24 + 1 x 30 + 1 x 28)/(1 + 1 + 1 + 1) = 27.0*

** mean.sq.devn = (1 x 1 + 1 x 9 + 1 x 9 + 1 x 1)/(1 + 1 + 1 + 1) = 5.0*

β = 0.18 RateRatio = exp(β) = 1.2										
w.mean	*20.4*	*27.0*	*26.8*	*26.3*	*27.9**	*28.7*	*28.0*	*26.0*	*21.3*	*23.7*
w.mean.sq.devn.	*8.4*	*14.9*	*7.5*	*3.9*	*4.5**	*9.8*	*6.3*	*10.3*	*8.1*	*11.1*

** mean = (1.44 x 26 + 1 x 24 + 2.99 x 30 + 2.07 x 28)/(1.44 + 1 + 2.99 + 2.07) = 28.0*

** mean.sq.devn = (1.44 x 3.5 + 1 x 15.1 + 2.99 x 4.5 + 2.07 x 0)/(1.44 + 1 + 2.99 + 2.07) = 4.5*

β = 0.34 RateRatio = exp(β) = 1.4										
w.mean	*21.4*	*28.7*	*27.6*	*26.8*	*28.5**	*30.1*	*28.6*	*27.4*	*22.4*	*25.0*
w.mean.sq.devn.	*4.7*	*7.4*	*4.2*	*3.2*	*3.5**	*7.2*	*2.8*	*7.3*	*5.6*	*6.5*

** mean = (1.96 x 26 + 1 x 24 + 7.53 x 30 + 3.84 x 28)/(1.96 + 1 + 7.53 + 3.84) = 28.0*

** mean.sq.devn = (1.96 x 6.2 + 1 x 20.2 + 7.53 x 2.3 + 3.84 x 0.2)/(1.96 + 1 + 7.53 + 3.84) = 3.5*

Figure 1*: A: the temperatures T (in ° Celsius, for the day-of-week-that-month 'strata' or 'matched sets' pertaining to the 10 events that occurred in a selected year. The provenance of these data is described in the Supplement. The temperature on the* ***day of the event*** *is indicated in* ***bold****. The asterisk in column 5 indicates that the calculations are presented in full for this selected column (matched set).*

B: Weighted Mean Square Deviation (MSD) calculations at selected values of β. Each `mean` *is a weighted average of the temperatures $T_1, T_2, \ldots T_{4/5}$ within a matched set, with weights* $\exp[\beta T_1]$, $\exp[\beta T_2]$, ... *or, equivalently, as shown, with re-scaled weights* $\exp[\beta T'_1]$, $\exp[\beta T'_2]$, ... *where* $T'_1, T'_2, \ldots T'_{4/5}$ *are measured relative to the minimum T in the riskset, Thus, the minimum temperature in the riskset has a weight of 1. Each* `mean.sq.devn` *is a weighted average of the squared deviations of* $T_1, T_2, \ldots T_{4/5}$ *from* `mean`, *using these same weights. (For the calculations involvingβ = 0, all values in the matched set receive the same weight). The detailed calculations are shown for the selected column (5*).*

Consider first the relatively narrow spread of *T*s in matched set 5, namely 24, 26, 28 and 30 (or 0, +2, +4 and +6 above the minimum in the matched set). With a Rate Ratio of 1.2 per degree C, the weights are $1.2^0 = 1$, $1.2^2 = 1.44$, $1.2^4 = 2.07$ and $1.2^6 = 2.99$, so that the weighted MSW is 4.5 (around a weighted mean of 27.9 C). At the other (more favourable) extreme, consider the more spread out T's of 18.5, 21.5, 25 and 30 in matched set 2 (or 0, +3, +6.5 and +11.5 above the minimum in the set): the weighted MSD in this set is 14.9. To be *conservative,* we might take the typical weighted MSD to be on the '*smaller'* side, say 4.5. Under this almost-worst case scenario, to achieve the SE of 0.075, equation(4) indicates we would need to study $\frac{1}{0.075^2} \times \frac{1}{4.5} =$ 40 events.

Number of cases to ensure a specified power

To plan for a given level of statistical power, the SE of $\hat{\beta}$ has to be envisioned under *two* scenarios, i.e., at the null, β_{null} (typically 0), and at the alternative, $\beta = \beta_{alt}$, so that they satisfy

$$Z_{\alpha/2} \times SE_{null} + Z_{\beta} \times SE_{alt} = \Delta, \qquad (5)$$

where $\Delta = \beta_{alt} - \beta_{null}$.

Suppose we wished to have 80% power against an alternative of β = 0.18. For planning purposes, we might use the data in Figure 1A as pilot data, and calculate (conservatively) that the typical weighted MSD per riskset will be 5 under the null and 4.5 under the alternative. Thus, the number of events, n, needs to satisfy the equation

$$\frac{1.96}{\sqrt{\text{MSW}_{null} \times n}} + \frac{0.84}{\sqrt{\text{MSW}_{alt} \times n}} = \Delta. \tag{6}$$

With our anticipated $\text{MSW}_{\text{null}} = 5$, and $\text{MSW}_{\text{alt}} = 4.5$, equation (6) indicates

$$n = \left(\left[\ 1.96 \div \sqrt{5.0} + 0.84 \div \sqrt{4.5}\ \right]\ \div 0.18 \right)^2 = 50\ events.$$

One notices from Figure 1B that the MSD (and thus the amount of information) per matched set diminishes rapidly the further β departs from the null. So, a *conservative n* is obtained by using the *non-null* information for *both* SE's. With these same error rates, and noting that 1.96+0.84 = 2.8, the equation simplifies to

$$\text{number of events} = \left(\left[\ 2.8 \div \sqrt{\text{NonNull MSD in Typical Riskset}}\ \right]\ \div \Delta \right)^2. \tag{7}$$

If we want an easier to remember (and again slightly conservative) formula, we can round 2.8^2 up to 8, to obtain

$$\text{number of events} = (8\ \div \text{NonNull MSD in Typical Riskset})\ \ \times (1/\Delta)^2\ . \tag{8}$$

In our example, with $(1/0.18)^2$ rounded up to 31, this comes out to (8/4.5) × 31 = 55 events.

Discussion

Intuitively, greater variation in the exposure makes it easier to detect/measure an exposure-response relationship. Since time-stratified case-crossover studies make comparisons *within* each time-matched set, the precision/power depends on the *within-matched-set* variation, and not on the *overall* variation in the exposure.[4] This 'local' variation can be much smaller: for example, in Figure 1A, while the MSD of the 44 temperatures around the overall mean of 23.8 is 19.4 C^2 (it would be even higher if it were derived from an even larger portion of the year), the typical within-matched-set MSD is only11.1 C^2. Thus, investigators need to pay attention to the expected within-set variation in exposures to ensure that an adequate number of cases is identified.

Unless the exposure-response relationship is quite strong, 'local' MSDs calculated at the null will suffice for planning purposes, since a sample size exercise is merely a rough projection of the likely precision/power. As the authors of a classic textbook[5] cautioned "There is usually little point in introducing fine detail into what are essentially rather crude calculations."

Investigators should not base them on implausibly large values of Δ, or think that any one study will settle the matter. Instead, they should consider how much information their study will contribute to a future meta-analysis. A former colleague of ours likened the question to how much to give when the collection plate is passed around in a house of worship: it is the *total* collected that matters in the end; in most such places, there is no 'requirement' for the size of an individual contribution.[6]

As we explain in the Supplemental Online Content, rather than present *separate* formulae for exposures measured on continuous and all-or-none exposures, we urge investigators to use the *common* principles involved. The Supplement also aims to demystify the calculations involved in the Maximum Likelihood estimation of the regression parameter, and the precision of the fitted coefficient.

References

1 Janes H, Sheppard L, Lumley T. Case–Crossover Analyses of Air Pollution Exposure Data: Referent Selection Strategies and Their Implications for Bias. *Epidemiology* 2005;16:717-726.

2 Hanley JA. Simple and multiple linear regression: sample size considerations. *Journal of Clinical Epidemiology* 2016;79:112-119.

3 McKeown-Eyssen GE, Thomas DC. Sample size determination in case-control studies: the influence of the distribution of exposure. *J Chronic Disease* 1985;38:559-568.

4 Künzli N, Schindler C. A call for reporting the relevant exposure term in air pollution case-crossover studies . *J Epidemiol Community Health* 2005;59:527-530.

5 Breslow NE, Day NE. Design Considerations. Chapter 7 in Statistical Methods in Cancer Research Volume II: The Design and Analysis of Cohort Studies 1987. IARC Scientific Publication No. 82.

6 Hernán MA. Causal analyses of existing databases: no power calculations required. *J Clinical Epidemiology* 2021: Aug 27;S0895-4356(21)00273-0. doi: 10.1016/j.jclinepi.2021.08.028.Online ahead of print.

SUPPLEMENTAL DIGITAL CONTENT

A WHERE DO THE FORMULAE COME FROM?

Sample size calculations are *pre-study* calculations that depend on the data-analysis method that will ultimately be used (i.e., post data-collection) Thus, to understand them, it is best to go through an actual data-analysis exercise, and to *anticipate* the results of the model-fitting. To this end we begin with a small dataset and to see close-up what aspects of the data determine the standard errors that emerge during the parameter-fitting. To keep the dataset small but real, we studied tornadoes, where the relationship between *T* and their rate is strong enough to 'see' in a study of just 10 cases. [To have the study design mimic a study of human events, we retain the matching on day-of-the-week and month]

Part A of eFigure 1 shows the *T'*s for each of 10 matched sets generated by the 10 tornadoes that occurred in the southern portion of a Canadian province during one selected year. Without loss of generality, we consider the temperature (*T*) on a specified day, rather than a lagged version of *T*, as the determinant of the expected event rate for such days. We limit ourselves to the same multiplicative model for event rates shown in equation (1) in the full text.

The SE of the **β** fitted by conditional logistic regression to the dataset in eFigure 1

The average of the 10 *T*'s on the 10 'event' days was 26.2°C, whereas, as is shown in the first row of part B, the average of the 10 column-specific averages was only 23.8°C. This indicates that the sign of the fitted gradient of the event rates over *T* will be positive.

A

	Thu May	Sun Jun	Sat Jun	Tue Jul	Wed Jul	Mon Jul	Sun Aug	Tue Aug	Thu Sep	Fri Sep		
	(1)	(2)	(3)	(4)	(5*)	(6)	(7)	(8)	(9)	(10)		
	15.0	25.0	28.0	26.5	26.0	26.5	29.0	20.5	**24.0**	20.0		
	23.0	18.5	20.0	24.0	24.0	27.5	**26.0**	23.5	19.5	**26.5**		
	13.5	**30.0**	29.0	**28.5**	**30.0**	26.5	20.0	23.5	15.0	16.0		
	20.0	21.5	25.5	26.0	28.0	21.5	29.5	**29.0**	20.0	22.5	Sum	Mean
	20.5		**22.5**	22.5		**32.0**					**262.0**	**26.2**

* *

B

β = 0
RateRatio = exp(β) = 1

	(1)	(2)	(3)	(4)	(5*)	(6)	(7)	(8)	(9)	(10)	Sum	Mean
w.mean	*18.4*	*23.8*	*25.0*	*25.5*	*27.0**	*26.8*	*26.1*	*24.1*	*19.6*	*21.2*	*237.6*	*23.8*
w.mean.sq.devn.	*12.7*	*18.3*	*11.3*	*4.3*	*5.0**	*11.2*	*14.3*	*9.4*	*10.2*	*14.6*	*111.3*	*11.1*

** mean = (1 x 26 + 1 x 24 + 1 x 30 + 1 x 28)/(1 + 1 + 1 + 1) = 27.0*
mean.sq.devn = (1 x 1 + 1 x 9 + 1 x 9 + 1 x 1)/(1 + 1 + 1 + 1) = 5.0

β = 0.1
RateRatio = exp(β) = 1.11

	(1)	(2)	(3)	(4)	(5*)	(6)	(7)	(8)	(9)	(10)	Sum	Mean
w.mean	*19.6*	*25.6*	*26.1*	*25.9*	*27.5**	*27.9*	*27.3*	*25.1*	*20.6*	*22.7*	*248.3*	*24.8*
w.mean.sq.devn.	*10.8*	*18.2*	*9.6*	*4.2*	*4.8**	*10.7*	*9.7*	*10.5*	*9.3*	*13.4*	*101.1*	*10.1*

** mean = (1.22 x 26 + 1 x 24 + 1.82 x 30 + 1.49 x 28)/(1.22 + 1 + 1.82 + 1.49) = 27.5*
mean.sq.devn = (1.22 x 2.2 + 1 x 12.2 + 1.82 x 6.3 + 1.49 x 0.3)/(1.22 + 1 + 1.82 + 1.49) = 4.8

β = 0.261
RateRatio = exp(β) = 1.3

	(1)	(2)	(3)	(4)	(5*)	(6)	(7)	(8)	(9)	(10)	Sum	Mean
w.mean	*21.0*	*28.0*	*27.3*	*26.6*	*28.2**	*29.5*	*28.4*	*26.8*	*21.9*	*24.5*	*262.0*	*26.2*
w.mean.sq.devn.	*6.3*	*10.9*	*5.7*	*3.6*	*4.0**	*8.6*	*4.1*	*9.0*	*6.8*	*8.6*	*67.6*	*6.8*

** mean = (1.69 x 26 + 1 x 24 + 4.79 x 30 + 2.84 x 28)/(1.69 + 1 + 4.79 + 2.84) = 28.2*
mean.sq.devn = (1.69 x 4.9 + 1 x 17.8 + 4.79 x 3.2 + 2.84 x 0)/(1.69 + 1 + 4.79 + 2.84) = 4.0

	(1)	(2)	(3)	(4)	(5*)	(6)	(7)	(8)	(9)	(10)	Sum	Mean
residual	***−7.5***	***2.0***	***−4.8***	***1.9***	***1.8***	***2.5***	***−2.4***	***2.2***	***2.1***	***2.0***	*0*	*0*

***eFigure 1**: A: the temperatures,(T, in ° Celsius) for the day-of-week-that-month 'strata' or 'matched sets' containing the 10 events that occurred in a selected year. The temperature on the **day of the event** is indicated in **bold**. The asterisk in column 5 indicates that the ML calculations are presented in full for this selected column.*

B: the calculations used in the pursuit of the Maximum Likelihood (ML) estimate of β, starting with the null value. Each `mean` *is a weighted average of the temperatures $T_1, T_2, \ldots\ T_{4/5}$ within a stratum, with weights* $\exp[\beta T_1]$, $\exp[\beta T_2]$, *... or, equivalently, as shown, with re-scaled weights* $\exp[\beta T'_1]$, $\exp[\beta T'_2]$, *... where $T'_1, T'_2, \ldots\ T'_{4/5}$ are measured relative to the minimum T in the stratum, Thus, the minimum temperature in the stratum has a weight of 1. Each* `mean.sq.devn` *is a weighted average of the squared deviations of $T_1, T_2, \ldots\ T_{4/5}$ from* `mean`*, using these same weights. (For the calculations involvingβ = 0, all values in the matched set receive the same weight). The detailed calculations are shown for the selected column (5*). The sum/mean at the right is the sum/mean over the 10 instances/cases. The ML iterations continue until the sum/mean of the 10 fitted/weighted means equals (balances) the sum/mean of the 10 (observed) temperatures on the days the events occurred.*

As we show in section B, the Maximum Likelihood value of β (0.261) is found by starting at β = 0 and proceeding, by a directed search, until one reaches a β value for which the sum of the *T*'s on the 10 days when the event occurred (the '*observed*' Ts) *equals* the sum of the '*fitted*' *T*'s on these 10 days. More important is the formula for its standard error, $SE[\,\hat{\beta}\,]$ = 0.1216. As we explain in the supplement, the SE is found by simply summing the MSD's in the

10 matched sets to arrive at a total of 67.6, and taking the square root of the reciprocal of this, i.e., $(1 / 67.6)^{1/2} = 0.1216$. Note, however, that the 10 matched-set-specific MSD's are not the 12.7, 18.3, … 14.6 in the first set of calculations in eFigure 1B. Since $\beta = 0$, these 'initial' MSD's are calculated by weighing the T's within the set equally; they sum to 111.3. The MSDs that sum to 67.6 were calculated as *weighted* MSD's, where the weights for the T's in each matched set are the rate ratios implied by the value of β and the T's in the set. The calculation of the weighted MSD is illustrated for matched set 5. In it, when $\beta = 0.261,$ the weights for the 4 *T*s of 24, 26, 28 and 30 (or *T*s of 0, +2, +4 and +6 above the minimum in the set) are exp(0 × 0.261) = $\underline{1}$, exp(2 × 0.261) = $\underline{1.69}$, exp(4 × 0.261) = $\underline{2.84}$ and exp(6 × 0.261) = $\underline{4.79}$ respectively. The MSD for matched set 5 was 4.0; the MSDs for the 9 other matched sets ranged from 3.6 to 10.9, and the typical MSD was 6.8.

It can be seen from eFigure 1B (and also for Figure 1 in the main text) that the further β is from 0, the smaller is the typical weighted MSD, and thus the larger is the SE of the fitted β: Whereas the SE is $(1 / 67.6)^{1/2} = 0.12$ at the ML value of $\beta = 0.261$, it is $(1 / 111.3)^{1/2} = 0.09$ at the null value of β. Thus, when β is further from zero, the required sample sizes will be larger than those illustrated in the earlier sections.

B MAXIMUM LIKELIHOOD ESTIMATION DEMYSTIFIED

Multinomial probabilities: the possible days an event could have occurred

As Chapters 13 and 15 of Clayton and Hills[1] show, and as Armstrong[2] re-iterates, the rate ratio in a person-time analysis of a binary exposure can sometimes be estimated by treating the total number of events within the stratum as a fixed quantity rather than the random variable that it is.

In the examples addressed in these chapters, how the events distribute themselves within the 'exposed' and 'unexposed' person-time can be described by a *binomial* random variable, in which the number of events serves as the '*n*' and the probability parameter is a function of the amounts of person time and the rate ratio. Parameter estimation is usually via Maximum Likelihood (ML). In *our* context, where the possible event days are days within a matched set, how the events distribute themselves over the possible days can be described by a *multinomial* random variable, in which the number of events (typically 1 per matched set) serves as the '*n*' and the multinomial probabilty parameters are a function of the temperatures and the rate ratios. For example, in column (5*) in eFigure 1, the temperatures on the 4 candidate days are 26, 24, 30 and 28 °C. Thus, *given* that an event occurred on one of these days (i.e., *conditional* on the event having occurred within the stratum), the multinomial probabilities that it occurred on the first, second, third or fourth of these days are, respectively,

$$\frac{\{\,exp[26\beta],\;\; exp[24\beta],\;\; exp[30\beta],\;\; exp[28\beta]\,\}}{exp[26\beta] + exp[24\beta] + exp[30\beta] + exp[28\beta]}\,.$$

These probabilities have the same structure as the probabilities that each of the nominees will win the Oscar[4-6] or the economic choices made by a consumer.[7-8]

The ML procedure for multinomial/conditional logistic regression, from first principles

The Method of Least Squares seeks the parameter value that minimizes the sum/average of the squared distances between the observed and fitted responses (the 'y's). Thus, since the quantity being 'optimized' uses the scale the responses are measured in, it is easily understood: if, for example, we fit a sine curve to the pattern of temperatures over the year, the goodness of fit criterion involves discrepancies in the °C scale. Very differently, *the Method of Maximum*

Likelihood seeks the parameter value that maximizes the sum/average of the logs of the probabilities of obtaining the data patterns that were observed. While the ML *principle* may be a natural one, the *scale* in which the criterion is measured is not so familiar. Nevertheless, as we will now see, the 'balancing equation' that must be satisfied/solved numerically is quite natural, even if it is not always emphasized.

To see why, we return to the data in column/stratum (5) in eFigure 1A, where the temperatures on the 4 candidate days are 26, 24, 30 and 28 °C and, thus, the multinomial probabilities that the event occurred on the first, second, third or fourth of these days are, respectively,

$$\frac{\{\,exp[26\beta],\ \ exp[24\beta],\ \ exp[30\beta],\ \ exp[28\beta]\,\}}{exp[26\beta]\ +\ exp[24\beta]\ +\ exp[30\beta]\ +\ exp[28\beta]}$$

The event occurred on the day when the temperature was 30 °C, and so the probability that it would have happened *on that day (rather than on one of the other three days)* is

$$\frac{exp[30\beta]}{exp[26\beta]\ +\ exp[24\beta]\ +\ exp[30\beta]\ +\ exp[28\beta]}$$

Thus, the log-likelihood contribution from this matched-set, i.e., the log of this probability as a function of β, is

$$30\beta\ - \log(exp[26\beta]\ +\ exp[24\beta]\ +\ exp[30\beta]\ +\ exp[28\beta])$$

The full log-likelihood is the sum, over the 10 matched sets, of the set-specific contributions. To maximize it with respect to β, one finds the value at which its derivative equals zero. For the log-likelihood contribution from matched set (5*), the derivative with respect to β is

$$30\ -\frac{exp[26\beta]\ \times 26\ +\ exp[24\beta]\ \times 24 +\ exp[30\beta]\ \times 30 +\ exp[28\beta]\ \times\ 28}{exp[26\beta]\qquad +\ exp[24\beta]\qquad +\ exp[30\beta]\qquad +\ exp[28\beta]}.$$

Although it may seem formidable, the quantity to the right of the minus sign is simply a *weighted mean of the 4 temperatures*, with weights given by the 4 exponentiated quantities. These weights are more manageable if we divide each of them by $exp[24\beta]$, so that the *lowest temperature in the matched set receives a weight of 1*, and so that the derivative (sometimes called the 'score') becomes

$$30\ -\frac{exp[2\beta]\ \times 26\ +\ 1\ \times 24 +\ exp[6\beta]\ \times 30 +\ exp[4\beta]\ \times\ 28}{exp[2\beta]\qquad +\ 1\qquad +\ exp[6\beta]\qquad +\ exp[4\beta]}.\ \ (S1)$$

We can think of the quantity after the minus sign as the "fitted" or "expected" value of the temperature on the day of the event, and thus we can rewrite the equation in which the derivative is set to zero (often called the 'estimating equation') as the 'balancing equation'

Sum(Observed *T* on day of event) = Sum(Fitted *T* on day of event),

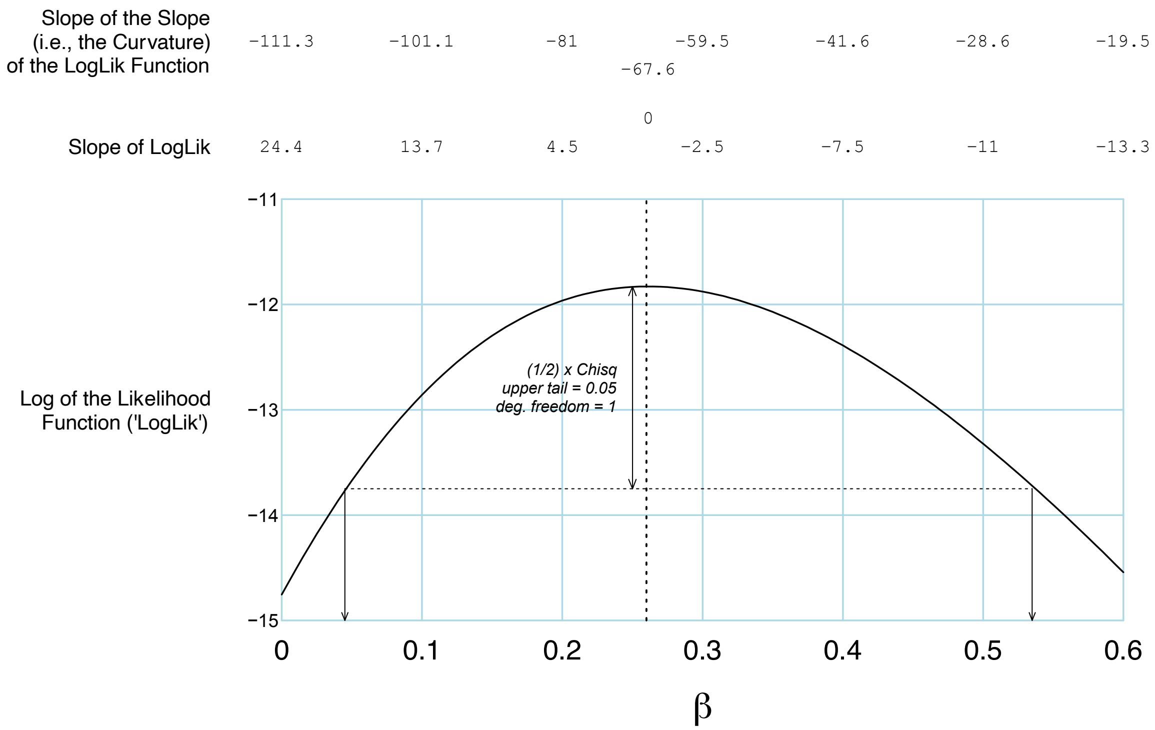
where the Sum is over the 10 matched sets.

Today, unlike when this model was first fitted in the mid 1960s, the search for the ML estimate can be easily carried out by trial and error using just a spreadsheet. As is shown at the right of eFigure 1, the sum / mean of the observed temperatures on the 10 'event' days is 262 / 26.2 °C. If there were no linear relation with *T*, i.e., if β = 0, then the (null) fitted sum / mean would be 237.6 / 23.8 °C. Since 26.2 is larger than expected, we need to 'move up' β until the fitted sum / mean equals the observed value. As can see seen in eFigure 1B, this 'balance' is achieved at β = 0.261.

In Least Squares regression, the "*y*" residuals must balance each other. Perhaps not so surprisingly, since *conditioning* reverses the $x \rightarrow y$ focus, in conditional logistic regression it is the residuals of the "*x*" values (the predictors in the regression) that must be balanced. Those familiar with fitting proportional hazards models may even recognize each difference between the observed and fitted temperature for the day of the event (shown in the last row of eFigure 1B) as a "Schoenfeld" residual.

The Precision of the ML estimate of the exposure-response parameter

Before statistical packages were readily accessible, a first course in simple linear regression usually introduced the closed-form formula for the standard error of the fitted slope. Very often however, it was shown in a form that involved the fewest computational steps rather than for illumination, and so opportunities to gain some intuition as to what determines the precision were lost.[9] This lack of transparency is even greater in the case of parameters fitted by ML, since the standard error is model-based, and calculated only after the solution (often iterative) is reached. Thus, in the didactic spirit of this note, we will show how the standard error output by a conditional logistic regression routine is easily calculated from a mere spreadsheet. Since our conditional logistic regression model involves just 1 parameter, the 'matrix inversion' that is a feature of most regression fits takes the simple form of $1/I$, where I is a scalar (1-dimensional) quantity. The reason for the choice of the letter I will become apparent later, and the 'I' quantity will play a central role in sample size projections.



eFigure 2*:Log-likelihood function for the parameter β of the exposure-response model, based on the data from the 10 matched sets in eFigure 1, together with its first and second derivatives computed at selected parameter values. The log-likelihood function reaches its maximum at β = 0.261, where its first derivative equals 0. The quantity 67.6 measures how curved the curve is at this ML value, and the square root of its reciprocal provides the Standard Error of the fitted β. The SE can then be used to form a Gaussian-based CI, or one can use the Likelihood ratio and the Chi-Square distribution to find the range of parameter values compatible with the data (limits are marked by the 2 arrows at 0.05 and 0.53).*

Before we introduce the formula-based approach that reveals where the precision (SE = 0.12) of the point estimate (0.261) comes from, we first use ‘brute force’ to numerically compute the standard error directly from the generic log-likelihood form. In other words, we rely *solely* on the log-likelihood *function* (‘LogLik’) plotted in eFigure 2. The ML estimate is the parameter value at which the first derivative (slope) of the log-likelihood function crosses from positive (at the left of the maximum) to negative (at the right), namely 0.261. Its variance is the reciprocal (inverse) of the (negative of the) second derivative of log-likelihood function evaluated at this same parameter value. This makes intuitive sense: the more concentrated (the sharper, or more curved) the curve is at its maximum, the narrower is the range of parameter values supported by the data. Moreover, as we go from left to right along the β scale, the log-likelihood curve goes from low to high to low, so its *slope* (the first derivative) goes from positive to negative, and so its *curvature* (the second derivative) is negative. *The more negative the curvature is, the tighter the log-likelihood and the more precise is the point estimate*.

One can check manually/visually that the first derivative is 4.5 at β = 0.2 and -2.5 at β = 0.3. Thus, the second derivative at β = 0.25 is approximately (-2.5 – 4.5)/0.1 or -70, and so its value of -67.6 at the ML value of β = 0.261 makes sense. R.A. Fisher, who developed the ML theory in the 1920s, called the -(-67.6) = 67.6 the '*Information'* (*I*) *in the data concerning β,* and showed that *its reciprocal (i.e., 1/I = 1/67.6) can be taken as the variance of the β estimate*, so that the Standard Error, the square root of the variance, is

$$\mathrm{SE}\left[\,\hat{\beta}\,\right] = (1/\mathrm{Information})^{1/2}\ ,$$

or, in this example,

$$SE\left[\,\hat{\beta}\,\right] = (1\ /\ 67.6)^{\,1/2}\ = 0.1216.$$

One can verify that this agrees with the output from the **`clogit`** function in R or Stata or the **`phreg`** (with the **`strata`** statement) procedure in SAS.

Even though most textbooks begin their teaching of Maximum likelihood by defining the Likelihood as a *product* of probabilities, Fisher always began directly with the *log*-likelihood, so that it can be immediately written as a *sum* of the individual *log*-likelihood contributions, one from each 'datapoint'. Quite apart from making the sum a more manageable number, the log-version immediately emphasizes that each datapoint (or matched set in our example) *adds* to the information about the parameter of concern, that not all datapoints contribute equally, and that we can readily quantify, in a technical sense, exactly how much 'information' each one adds.

As we will now demonstrate, by working with this formal measure of information, and just taking the reciprocal of the combined information at the very end, the factors that determine the variance and the SE of the fitted β become very clear. So, instead of relying on the numerical version of the second derivative of the entire log-likelihood function as we did above, we will now show the specific *closed-form formula* that measures the 'information' contributed by each riskset, using as an example that contributed by riskset 5. From equation (S1) above giving the formula for the first derivative for the log-likelihood contribution, one can use the rules of calculus to verify that the second derivative involves the same weights used in the weighed mean of the 4 temperatures, and that it is merely the negative of the weighed MSD of these 4 temperatures from that matched-set-specific weighed mean. The calculation of this weighted mean square is illustrated for selected matched set (5) in eFigure 1, where it is calculated under 3 scenarios: the null and ML values of β, and an intermediate value where $\beta = 0.1$. The 4 temperatures are 26, 24, 30 and 28, or, (measured from their minimum), +2, 0, +6 and +4. Thus, at $\beta_{ML} = 0.261$, so that $\exp(\beta_{ML}) = 1.3$, the weights are $1.3^2 = 1.69$; $1.3^0 = 1$; $1.3^6 = 4.79$; and $1.3^4 = 2.84$, so the weighted mean is 28.2. The weighted mean of the squared deviations of the 4 temperatures from this 28.2 is 4.0. As such, matched set (5) is the one with the second-smallest spread of temperatures, and it contributes the second smallest amount of information to the combined information of $I = 67.6$. The smallest contribution of the 10 matched sets is the 3.6 from riskset (4), where the temperature range was just 4.5 °C, and the largest is the 10.9 from set (2), where the range was 11.5 °C. This ranking is the same as when the information is calculated at $\beta_{NULL} = 0$. That the SE of the fitted slope is inversely related to the spread of the exposure variable makes explicit what researchers instinctively know: it is difficult to measure a slope (e.g., the fuel consumption of a vehicle) over a short distance.[10]

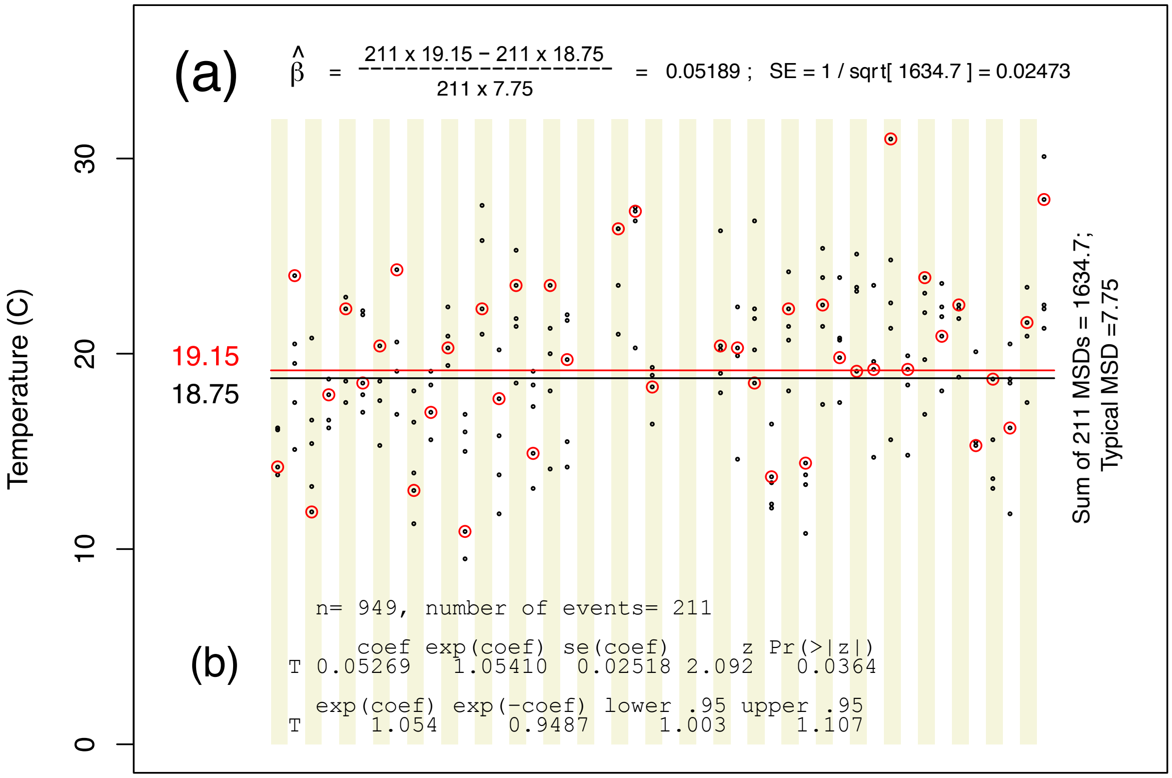
Readers may wonder why we do not refer to the weighed mean square deviation as a '*variance'*. Technically it is, but since most readers associate a variance with a divisor that is one less than the number of objects, we prefer to use the more expressive term mean square deviation. In his seminal article, Cox[11] refers to it as a "variance over the finite population of *T*'s using an 'exponentially weighed' form of sampling." This fits with the principle that in a regression model, that x's are not treated as realizations of a random variable whose variance is to be *estimated*;[9-10] the regressors are considered fixed, as if they had been decided by the investigator.

Fisher made a distinction between the *expected* information concerning β calculated using *pre-study projections* and the *observed* information calculated *post study* using the observed data. The latter is used to calculate the Standard Error for the β estimate, namely

$$\mathrm{SE}\left[\,\hat{\beta}\,\right] = (1 / \mathrm{I}\,)^{1/2} = (1 / [6.3 + 10.9 + \ldots\ 4.0 + \ldots + 6.8 + 8.6]\,)^{1/2} = (1 / 67.6)^{1/2} = 0.1216.$$

Smaller signal, more matched sets

To illustrate this, we extended our case series to the 211 events that occurred in the same Canadian province during the full 30-year period for which events were documented. To more easily distinguish the matched sets, the 4/5 datapoints shown in each column of eFigure 3 are the temperatures on the same day-of-week in the same-month for *every fifth on*e of these 211 matched sets. To dilute the relationship, we used temperatures from another Canadian province, and used mean temperature rather than maximum temperature.



```
  n= 949, number of events= 211

      coef exp(coef) se(coef)     z Pr(>|z|)
T 0.05269   1.05410  0.02518 2.092   0.0364

  exp(coef) exp(-coef) lower .95 upper .95
T     1.054     0.9487     1.003     1.107
```

eFigure 3*: The black dots are the temperatures (T's) for every fifth one of the 211 matched sets (see text). The temperature on the day of the event is indicated by a red circle. The 19.15 on the left hand side is the mean of the temperatures for the 211 event days, while the 18.75 is the mean of the 211 matched-set means. The typical MSD is 7.75 (right hand side), and the sum of the 211 MSDs is 1634.7. The first approximation to the parameter of interest, along with its standard error (SE), is shown is shown in (a), while the ML parameter estimate and its SE (fitted via conditional logistic regression,* `clogit` *in* R*) are shown in (b).*

The fitted β is now 0.053, but because the SE is 0.025, the z statistic is just over 2, and very similar to the z statistic of 0.261/0.122 in our earlier example with just 10 events but a stronger signal. The greater precision is largely because of the larger number of events (211). Since the fitted β is much closer to zero, the typical MSD at the ML value (7.47) is very close to the 7.75 calculated at β = 0. Thus, the SE of 1/sqrt(7.47 × 211) = 0.0251 is only very slightly larger than the SE of 1/sqrt(1634.7) = 0.0247 calculated at the null.

Those who prefer to stay close to their data can avoid the conditional logistic regression software altogether when β is expected to be very close to 0. The first iteration of the ML procedure has the simple form shown in expression (a) in eFigure 3, and yields a very good approximation to the deluxe final ML version. This very good closed form approximation in the case of weak relationships is not well known, although it was mentioned in the report[12] of a well-chronicled study of the health effects of environmental contamination.[13-14] That 1986 study had the same matched-set structure as the illustrations used here.

C A UNIFIED APPROACH TO ALL-OR-NONE AND QUANTITATIVE EXPOSURES

In our calculations thus far, there was nothing special about the fact that T is recorded on an interval scale. *Had the exposure been recorded on an all-or-none (2-point, binary) scale, the approach would have been exactly the same*: the only change would be the focus on a single Rate Ratio = exp[β] contrasting the rates in the presence and absence of the factor of interest and the 0/1 exposure scale in which the (weighted) means and squared deviations are measured. To make these ideas more concrete, eFigure 4 revisits the 10 events in eFigure 1, but shows a binary exposure (to stay with the same illustrative example, we merely dichotomized the temperature scale.)

A

	Tue Aug (1)	Tue Aug (2)	Tue Aug (3)	Tue Aug (4)	Tue Aug (5*)	Tue Aug (6)	Sun Aug (7)	Sun Aug (8)	Sun Aug (9)	Sun Aug (10)		
	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	**0.0**	0.0		
	0.0	0.0	0.0	0.0	0.0	1.0	**0.0**	0.0	0.0	**0.0**		
	0.0	**1.0**	1.0	**1.0**	**1.0**	0.0	0.0	0.0	0.0	0.0		
	0.0	0.0	0.0	0.0	1.0	0.0	1.0	**1.0**	0.0	0.0	Sum	Mean
	0.0		**0.0**	0.0		**1.0**					**5**	**0.5**

* *

B

β = 0
RateRatio = exp(β) = 1

	(1)	(2)	(3)	(4)	(5*)	(6)	(7)	(8)	(9)	(10)	Sum	Mean
w.mean	*0.00*	*0.25*	*0.40*	*0.20*	*0.50**	*0.40*	*0.50*	*0.25*	*0.00*	*0.00*	*2.50*	*0.250*
w.mean.sq.devn.	*0.00*	*0.19*	*0.24*	*0.16*	*0.25**	*0.24*	*0.25*	*0.19*	*0.00*	*0.00*	*1.51*	*0.152*

** mean = (1 x 26 + 1 x 24 + 1 x 30 + 1 x 28)/(1 + 1 + 1 + 1) = 0.50*
mean.sq.devn = (1 x 1 + 1 x 9 + 1 x 9 + 1 x 1)/(1 + 1 + 1 + 1) = 0.25

β = 0.8
RateRatio = exp(β) = 2.23

	(1)	(2)	(3)	(4)	(5*)	(6)	(7)	(8)	(9)	(10)	Sum	Mean
w.mean	*0.00*	*0.43*	*0.60*	*0.36*	*0.69**	*0.60*	*0.69*	*0.43*	*0.00*	*0.00*	*3.78*	*0.378*
w.mean.sq.devn.	*0.00*	*0.24*	*0.24*	*0.23*	*0.21**	*0.24*	*0.21*	*0.24*	*0.00*	*0.00*	*1.63*	*0.163*

** mean = (4.95 x 26 + 1 x 24 + 121.51 x 30 + 24.53 x 28)/(4.95 + 1 + 121.51 + 24.53) = 0.70*
mean.sq.devn = (4.95 x 12.3 + 1 x 30.3 + 121.51 x 0.2 + 24.53 x 2.3)/(4.95 + 1 + 121.51 + 24.53) = 0.21

β = 1.6
RateRatio = exp(β) = 4.95

	(1)	(2)	(3)	(4)	(5*)	(6)	(7)	(8)	(9)	(10)	Sum	Mean
w.mean	*0.00*	*0.62*	*0.77*	*0.55*	*0.83**	*0.77*	*0.83*	*0.62*	*0.00*	*0.00*	*5.00*	*0.500*
w.mean.sq.devn.	*0.00*	*0.23*	*0.18*	*0.25*	*0.14**	*0.18*	*0.14*	*0.23*	*0.00*	*0.00*	*1.35*	*0.135*

** mean = (24.53 x 26 + 1 x 24 + 14764.78 x 30 + 601.85 x 28)/(24.53 + 1 + 14764.78 + 601.85) = 0.80*
mean.sq.devn = (24.53 x 15.3 + 1 x 35 + 14764.78 x 0 + 601.85 x 3.7)/(24.53 + 1 + 14764.78 + 601.85) = 0.14

	(1)	(2)	(3)	(4)	(5*)	(6)	(7)	(8)	(9)	(10)	Sum	Mean
residual	***0.0***	***0.4***	***−0.8***	***0.4***	***0.2***	***0.2***	***−0.8***	***0.4***	***0.0***	***0.0***	*0*	*0*

eFigure 4*: A: Exposures, recorded on a 0/1 scale, for the day-of-week-that-month 'column' containing each of the 10 events shown in Figure 1. The exposure on the day of the event is indicated in bold. The 2 columns in which there is no variation in exposure are non-contributory. In the remaining 8, the exposure factor was present on 5 of the days when the event occurred, and was absent on 3.*

B: The calculations used in the pursuit of the Maximum Likelihood (ML) estimate of β, exactly as in eFigure 1, beginning with β = 0, and ending with β = β_{ML} = 1.6. The SE for the fitted β is 1/sqrt[1.35] =0.86.

In the first two rows of eFigure 4B, readers can note one major simplification: in columns (3), (4) and (5), where the (unweighted) means of the 0s and 1s are 0.4, 0.2, and 0.5, the respective mean square deviations are the '*Bernoulli'* variances, 0.4 × 0.6 = 0.24, 0.2 × 0.8 = 0.16 and 0.5 × 0.5 = 0.25. However, they are not necessarily smaller when the weights are calculated at non-null values of β. The maximum information that any matched can provide is 0.5 × 0.5 = 0.25; this occurs when the exposure factor is equally likely to be present/absent, and the information is calculated at the null. Further away from these situations, the contribution per matched set can be less. Thus, in equation (6) in the fulltext the divisors of the 1.96 and 0.84 will be much smaller than they would with a quantitative *T* scale (of course, in the case of a truly binary exposure, the Δ of concern would likely be larger than the 0.1 employed there).

Using the Bernoulli (and weighted Bernoulli) variances in formula A, one arrives at the same sample size suggestions as those given by the specialized packages or tables. As an example, suppose we wished to have 80% power against an alternative of β = 0.8. Again. we could use the data in eFigure 1 as pilot data, and calculate that the typical information per matched set is 0.15 under the null and 0.16 under the alternative. Thus, the suggested number of events, *n*, is $([1.96 / \text{sqrt}(0.15) + 0.84 / \text{sqrt}(0.16)] / 0.8)^2 = 80$. Using an exposure prevalence of 0.25, a Rate Ratio of exp(0.8) = 2.25, interpolation between rows 3 and 23, and columns 2 and 3, in Table 7.9 of Breslow and Day (1987), and scaling up slightly to have the average of 4.4 (rather than 1 + 4 = 5) observations per matched set, yields an *n* of 82.

eFigure 4 shows why smaller matched sets involving a binary exposure are more likely to lack exposure variation and thus to be uninformative. A wider range of ‘days’ may reduce the effects of autocorrelation and increase the variation but may involve a bias/precision tradeoff.

In Table 2 of Lagakos et al.[12] the17 risksets ranged in size from 84 to 290. The proportions exposed varied from 0.18 to 0.40, so all risksets were informative. Interestingly, the authors approximated the ML parameter estimate using the closed form shown in eFigure 3(a) and intimated that the resulting β value of 1.11 may be an underestimate; in fact, the ML estimate is 0.99, and the SE at this value is smaller than it was at the null. Thus, when large effects of a binary exposure are anticipated, calculations at the null and at the alternative can be helpful to see how much information each matched set may contribute.

D THE PRICE OF ADJUSTING FOR OTHER VARIABLES

The *n*'s in the various formulae in the fulltext can be corrected upwards to allow for the loss in precision that occurs when one has to adjust for confounding variables. The loss is readily understood in the context of an occupational epidemiology example (8): imagine one wishes to quantify the effect of each year worked in a noisy workplace on hearing loss, while taking care to remove from the crude estimate the (also substantial) effect of age. If one pays no attention to age when selecting workers, one is likely to end up with a sample where age and duration of work are highly correlated, so that the joint distribution of work duration and age looks like an tilted ellipse or diamond or the like. Within each age-slice (i.e., when 'adjusting' for age) there is a narrow spread of numbers of years worked, and this a less precisely estimated slope. Suppose instead one selects workers from each of several age slices, not randomly, but on the basis of years worked, so that, within each age-slice, there is the widest possible spread of numbers of years worked. This creates a greater degree of ''balance'' (a lower correlation) between age and work duration, and the joint distribution of work duration and age is more orthogonal. Now, within each age-slice (i.e., when adjusting for age) there is a full spread of numbers of years worked, and this a more precisely estimated slope. [Similar reasoning applies when trying to separate the effect of egg yolk consumption on carotid plaque from the effect of bacon consumption (15)].

The same issue arises in trying to separate the effect of temperature and humidity on event rates. If the two are closely correlated, then when one conditions on (holds constant) humidity, the effective range of temperature is much reduced, and this makes the estimate of the net effect of temperature much less precise.

As we explain elsewhere (9), the variance inflation (VI) can be measured mathematically as the ratio of the area (volume) of the untitled region to the area(volume) of the tilted region containing the exposure and confounding variable(s). This ratio is none other than the reciprocal of the complement of the square of the simple/multiple correlation of the exposure of interest (E) with the confounder(s) (C) being adjusted for. Thus, the upwards corrected number of cases *n** becomes

$$n^* = \frac{n}{1 - (r_{E\,with\,C})^2},$$

where *n* is the number derived from the (E only) formula in the fulltext, and *r* is the anticipated multiple correlation coefficient between E and the confounding variable(s) C.

References

1 Clayton D, Hills M. Statistical models in epidemiology. Oxford University Press. New York. 1993.

2 Armstrong BG, Gasparrini A, Tobias A. Conditional Poisson models: a flexible alternative to conditional logistic case cross-over analysis. *BMC Medical Research Methodology* 2014, 14:122 http://www.biomedcentral.com/1471-2288/14/122

3 Cox DR. Regression models and life-tables. *J Royal Statistical Society, Series B*, 1972;34:187-220.

4 Pardoe, I. Just how predictable are the Oscars? *Chance* 2005;18:32–39.

5 Pardoe I. Predicting Oscar Winners. *Significance* 2007;4:168-173.

6 Pardoe I, Simonton DK. Applying discrete choice models to predict Academy Award winners*. J. R. Statist. Soc. A.* 2008;171:375–394.

7 McFadden, D. Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Frontiers in Econometrics*, 1974:105–142. New York: Academic Press.

8 McFadden D. Nobel Lecture. 2000 https://www.nobelprize.org/uploads/2018/06/mcfadden-lecture.pdf

9 Hanley JA and Moodie EEM. Sample Size, Precision and Power Calculations: A Unified Approach . J Biomet Biostat 2011; 2-5. http://www.medicine.mcgill.ca/epidemiology/hanley/Reprints/UniversalSampleSize.pdf

10 Hanley JA. Simple and multiple linear regression: sample size considerations. *Journal of Clinical Epidemiology* 79 (2016;79:112-119.

11 Cox DR. Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B* 1972; 34:187-220.

12 Lagakos SW, Wessen BJ, Zelen M. An Analysis of Contaminated Well Water and Health Effects in Woburn, Massachusetts. *Journal of the American Statistical Association* 1986;395:583-596.

13 Harr J. A Civil Action, Random House, New York. 1995.

14 Zaillian S. (Director). (1998). *A Civil Action* [Film]. https://www.imdb.com/title/tt0120633/

15 Spence JD, Jenkins JA, Davignon J. Egg yolk consumption and carotid plaque. *Atherosclerosis* 224:469-473.

Outlook

EE Decision

From em.ee.0.7dd016.6ac48f16@editorialmanager.com
<em.ee.0.7dd016.6ac48f16@editorialmanager.com>
on behalf of
Environmental Epidemiology <em@editorialmanager.com>

Date Mon 2022-09-05 6:40 AM

To James Hanley, Dr. <james.hanley@mcgill.ca>

Sep 05, 2022

RE: EE-D-22-00028, entitled "Planning and understanding sample sizes for case-crossover studies of environmental exposures"

Dear Dr. Hanley,

The editorial and peer review of your manuscript have now been completed. I regret to inform you that Environmental Epidemiology is unable to publish your manuscript.

The reviewer's comments are included at the bottom of this letter. Please note that the comments are for your information only and do not constitute a request for revisions.

Environmental Epidemiology hereby releases your manuscript from consideration so you may submit it elsewhere.

Thank you for submitting your manuscript to the Environmental Epidemiology.

Sincerely,

Prof Bert Brunekreef
Editor in Chief
Environmental Epidemiology
https://www.editorialmanager.com/ee/

Your username is: JHanley-733
https://www.editorialmanager.com/ee/l.asp?i=40014&l=KSWDTZSC

Reviewer Comments:

Reviewer #1: The manuscript describes power and sample size computation for studies applying the case-crossover design, together with algebraic definitions. The topic is of interest, with compelling discussions on several aspects and an illustration of real-data applications. However, in the current form, the manuscript is very hard to follow, with superficial information provided in the main text, a lack of description of the basic design settings, and confusion due to the use of different data examples. Detailed comments are provided below.

1. I found the structure chosen for this contribution very confusing. First, the authors provide limited information in the manuscript about the design per se (see Comments 7-9 below) and the statistical quantities (see Comments 11-12), with most of the latter confined to the appendix. Second, the authors use two different data examples, one very basic for the main paper and one more structured and coming from a real dataset in the appendix. Third, the appendix includes a long description of the underpinning likelihood theory, which, while interesting, is not essential. These issues result in a very superficial description being offered just by reading the article, which is not well complemented by the very complex theoretical overview provided in the appendix. I suggest the authors consider revising the article using a different structure and content.
2. One very simple way of doing it would be to use a single real-data example (I would recommend the tornado data), and illustrate all the examples and steps using it. For instance, the results of the equations for the minimum sample size and power can be computed for this (relatively) small dataset, as well as the MSD/MSW and the likelihood contribution for each risk set.
3. Similarly, the algebraic definitions of the MSD/MSW should be provided in the main text, as they are central to the power calculations.
4. I am not sure all the information provided in the appendix is relevant. For instance, most of the likelihood theory (part B) is interesting, but very general and not specific to this context. If the authors choose to keep it, I would move it to the very end of the appendix and focus first on more specific aspects related to the case-crossover design.
5. The provision of a code, for instance using the R software, would enormously facilitate the application of these methodologies by the users. I suggest providing a script and real-data example, possibly replicating the results described in the manuscript.
6. The authors chose to focus their contribution on the application of the case-crossover design in environmental research. However, exactly the same models and power calculations for this design can be used in other epidemiological areas. I wonder if the authors can make the presentation more general, and use the application for studying risks associated with environmental exposures only as an example. This would make the contribution relevant to a broader audience.
7. The authors need to contextualise the use of power calculations in this setting. In the majority of the cases, the data collection and analysis of a case-crossover study do not require the drafting of a pre-specified protocol and are easy to perform. This does not mean that power calculation is not required, but the authors should clarify at which step of a project a researcher can find it useful.
8. The authors provide very little detail on the structure of the case-crossover design, and given the type of contribution (Education Corner), it cannot be assumed that all the readers are familiar with it. In particular, several control sampling schemes exist, and the authors only describe the most commonly applied in environmental epidemiology, i.e. the time-stratified with month/weekday strata. The authors should provide a more general presentation, describing the general structure of the design and motivating the decision about presenting this specific scheme.
9. There are very few references in the article. The authors can include up to 30 of them, and I strongly suggest adding more.
10. For the description of the method, if the authors choose to follow my suggestion in Comment 1, I would refer to an 'exposure of interest x' rather than temperature or any specific factor. This can make the illustration more general. The authors can then add a specific example using real data.

11. The description of the statistical model and the quantities of interest are confused. For example, in Equation 1 the outcome lambda does not represent a rate (e.g., number of events per day), but an individual hazard, as in Cox proportional hazard model. The case-crossover design works with individual data. It is true that with aggregated data its likelihood can be replicated using conditional Poisson models (see reference 2 in the appendix), which work with rates. However, in this case, the analysis is not performed using a logistic regression, which only accepts a binary response and not a count. I suggest the authors describe the individual-level setting and then mention the case of aggregated data as a specific extension.
12. The authors should make clearer to the reader what are the theoretical foundation of the separation in weaker and stronger-relationship scenarios. My understanding is that it is related to the possibility of making an approximation in the weaker-scenario setting, but is it unclear how the threshold of an RR. I could not find a clear explanation neither in the main manuscript or in the appendix.
13. There is another complication that is not addressed in this contribution. The typical case-crossover setting can be cast as a nested case-control in which a selection of potential control-days are taken in each risk set. For instance, in the time-stratified sampling scheme, the same weekdays in the months are selected, although it is possible to select all the other days in the month and to control by day of the week directly (a sort of full-stratum scheme). In this case, it can be expected that the latter method provides more power, as more controls are selected for each case. The authors can refer to power calculation in nested case-control studies for references. I assume that this is not clear in this contribution as the authors keep referring to an aggregated data structure, in which days and not individual events are the unit of analysis. In this case, days are repeatedly taken both as cases and controls, and therefore the problem above does not apply. However, for individual-level analyses, I assume that another parameter that influence power is the number of control taken in each risk set. The authors should at least mention this issue.
14. It would be good that the authors clearly state somewhere in the article that the estimators from logistic regression in case-crossover design define a log risk ratio, not a log odds ratio. This is implied in some parts when describing the method, but never clearly stated. I think this is not clear to the majority of users of the design.
15. I strongly suggest defining the multiplier for the confidence level explicitly, instead than approximating it to 2 for the 95% option. Similarly, the concepts of type-I and type-II errors should be described, and the related quantities (e.g., alpha and 1-alpha) defined. For instance, it is not clear to the reader what Z_beta=0.94 really is on page 9.
16. The 'margin of error' seems the width of the confidence interval. If so, I would refer to it explicitly.
17. Add numbers for all the equations.
18. The title is not appropriate. I would refer to 'power and sample size'.

Reviewer #2: This manuscript deals with providing a means of properly considering sample size and statistical power issues when dealing with time-stratified case-crossover studies. The focus is on studies that deal with assessment of acute (i.e., short term) health impacts of environmental exposures, but the issues discussed have the potential to inform studies in any other domain area with similar study designs.

Several important issues pertinent to sample size and statistical power aspects related to this unique, yet very important, study design are discussed. In particular, the need to pay attention to the proper study unit in determining sample size and power and also the unique importance of the within matched-set variation in exposures (which is expected to be relatively much small than the variance of exposures to be observed over the entire study period. Both of the above issues are important and have direct impact on determining sample sizes (and hence the eventual value of any given relevant study). The Maximum Likelihood based calculations for the proposed formulae, albeit ad-hoc and very informal, are based on parameters to be estimated under the proper conditional logistic regression modeling framework.

Major Issues

The biggest problem with the paper is the ad-hoc nature of the proposed techniques for sample sizes and statistical power, and the way they are justified.

The insights provided, while important and worthy of serious consideration, are not particularly novel. These are facts that ought to be well known and practiced by any properly trained and experienced biostatistician and epidemiologist involved in such studies.

As written, the manuscript reads like a primer for best practices in conducting such studies and as a precautionary document for pitfalls against ending up with inappropriately designed time stratified case-crossover studies.

The differences and nuances between the "weaker" and "stronger" relationship scenarios are not well developed and articulated.

The supplement provides more details, but they all remain to be summaries of well known principles, based on informal reasonings and calculations - even though several relevant issues are discussed such as issues related to the nature of the exposure (continuous vs. categorical) and issues with adjustments for other variables.

Reviewer #3: This is a very pedagogical paper that illustrates how to make simple size calculations for case-crossover studies. The article can be useful for planning purposes, as it provides examples of how to calculate the relevant parameters using a spreadsheet. Besides, in contains an interesting supplement that provides more insights.

Comments:

1) typo in page 6: 've'.

2) 'MSW' in page 9. Should it be 'MSD' instead? If not, MSW has not been defined.

3) A case-control study can be seen as a particular case of a time series study (https://pubmed.ncbi.nlm.nih.gov/16809430/). Thus, it may be interesting that the authors discuss the approach taken by Armstrong et al (https://doi.org/10.1186/s12874-019-0894-6). They discuss that it may be useful for case-crossover analysis. It would be good to describe the similarities, and even if one reaches the same conclusions trying to replicate one of their examples.

4) Supplement, page 17. The last sentence in section A can be misleading. Actually, as the authors mention elsewhere, when beta is further from zero we are trying to detect a stronger effect and

the required sample size will be smaller. What the authors really mean is that assuming a beta further from zero leads to larger values of the standard error. The overall effect of increasing beta on required sample size would be the combination of these two effects.

5) the symbol in eFigure 3 is not seen properly in my file.

6) Page 28: the authors mention that in this case the MSD is not necessarily smaller when weights are calculate at non-null vales of beta. However, in the main text (pages 9 and 10) it seems the authors suggest this is always the case. It would be good to clarify if this is always the case, or if it was just a trend observed in the example.

7) Page 29: the reference to eFigure 1 should be eFigure 4, I think.

8) It could be useful to provide a spreadsheet with the data used in the examples and the formulas implementing the calculations.