Case-Crossover Analysis in Air Pollution Epidemiology

Ho Kim
Seoul National University
School of Public Health

Case-Crossover Analysis

- Popular tool for estimating the effects of acute outcomes by environmental exposures.
- Only cases are sampled, estimates are based on within-subject comparisons of exposures at failure times vs. control times
- Controls for time-invariant confounders by design
- <u>Problems</u>: selection bias

    confounding by time-varying factors

    Time trends in exposure of interest -> bias

Reference (1)

- Maclure(1992) Am J Epi 133:144-153

  The case-crossover design : a method for studying transient effects on the risk of acute events
- Mittleman, Maclure, Bobinson(1995) Am J Epi 142:91-98

  Control sampling strategies for case-crossover studies :

  an assessment of relative efficiency
- Lee, Kim, Schwartz(2000) Environ Health Persp 108:1107-1115

  Bidirectional Case-crossover studies of air pollution:

  Bias from skewed and incomplete waves
- Bateson and Schwartz(2001) Epidemiology 12:654-661

  Selection bias and confounding in case-crossover analyses of environmental time-series data

Reference (2)

- Navidi & Weinhandl(2002) Epidemiology 13:100-105

  Risk set sampling for case-crossover design
- Lee, Schwartz(1999) Environ Health Persp 170:633-636

  Reanalysis of the effects of air pollution on daily mortality in Seoul, Korea: A case-crossover design
- Kwon, Cho, Nyberg, Pershagen(2001)

  Epidemiology 12:413-419

  Effects of ambient air pollution on daily mortality in a cohort of patients with congestive heart failure

Bidirectional Case-crossover Studies of Air Pollution: Bias from Skewed and Incomplete Waves

Lee, Kim, Schwartz (Environ Health Persp, 2000) 108:1107-1115

- Sampling selection strategy
  Unidirectional(retrospective,prospective),Bidirectional
  num.of controls(1,2)
- Exposure pattern Left, right skewed
  Cup of Cap shape
- Incompleteness
- Bidirectional is better than unidirectional.
- Bidirectional fails with incomplete exposure.

Table 3. Comparison of case- crossover estimators by various sampling approaches in a situation where both long-term time trends (decreasing with calendar time overall) and seasonal waves of SO_2 levels exist.

Estimator (true β = 0.001)[a]	Right heavier			Left heavier			Symmetric		
	Mean	SE ($\times 10^{-3}$)	RMSE ($\times 10^{-3}$)	Mean	SE ($\times 10^{-3}$)	RMSE ($\times 10^{-3}$)	Mean	SE ($\times 10^{-3}$)	RMSE ($\times 10^{-3}$)
β_{U1-}	–0.00012	0.8576	1.4074	–0.00173	0.7447	2.8308	–0.00009	0.7729	1.3369
β_{U2-}	–0.00057	0.8161	1.7723	–0.00324	0.7961	4.3131	–0.00083	0.6882	1.9554
β_{B2}	0.00090	0.5784	0.5978	0.00100	0.6535	0.6535	0.00099	0.5969	0.5869
β_{B4}	0.00090	0.4806	0.4912	0.00097	0.5810	0.5817	0.00087	0.5340	0.5504
β_{U1+}	0.00200	0.8562	1.3187	0.00374	0.7090	2.8291	0.00214	0.7428	1.3573
β_{U2+}	0.00255	0.8166	1.7554	0.00530	0.7282	4.3599	0.00293	0.6340	2.0279

RMSE, root mean-squared error.

[a] β_{U1-}, unidirectional retrospective with one control; β_{U2-}, unidirectional retrospective with two controls; β_{B2}: bidirectional with two controls; β_{B4}, bidirectional with four controls; β_{U1+}, unidirectional prospective with one control; β_{U2+}, unidirectional prospective with two controls.

Selection Bias and Confounding in Case-Crossover Analysis of Environmental Time-series Data

Bateson and Schwartz(Epidemiology, 2001) 12:654-661

- Simulation study of the sensitivity of the selection bias
- Selection bias results when exposure in the reference period is not identically representative of exposure in the hazard period (This bias can be estimated and removed)
- Confounding results from a common temporal pattern in the exposure and outcome time-series that are correlated in finite series length.
- All biases are reduced by choosing shorter referent-spacing length.

TABLE 1. Results Are the Coefficients of Effect of Particulate Matter <10 μm in Aerodynamic Diameter per 100 μg/m³ on Total Mortality in Cook County (1988–1993) from a Generalized Additive Poisson Regression Using 24 Degrees of Freedom to Control for Season and from Symmetric Bidirectional (SBI) Case-Crossover Analyses with Different Lag Lengths from 6 to 14 Days

Analytic Method	Spacing between the Hazard and Reference Days	Beta	Standard Error	95% CL
Poisson Regression		0.0559	0.0164	0.0238, 0.0880
SBI case-crossover	6 days	0.0477	0.0204	0.0076, 0.0877
SBI case-crossover	7 days	0.0462	0.0203	0.0064, 0.0860
SBI case-crossover	8 days	0.0383	0.0203	−0.0014, 0.0780
SBI case-crossover	9 days	0.0301	0.0201	−0.0093, 0.0695
SBI case-crossover	10 days	0.0501	0.0202	0.0105, 0.0896
SBI case-crossover	11 days	0.0660	0.0199	0.0270, 0.1050
SBI case-crossover	12 days	0.0661	0.0198	0.0272, 0.1049
SBI case-crossover	13 days	0.0738	0.0197	0.0352, 0.1123
SBI case-crossover	14 days	0.0732	0.0198	0.0344, 0.1121
SBI case-crossover	6–14 days Inclusive*	0.0576	0.0175	0.0233, 0.0920

All models controlled for temperature, barometric pressure, and relative humidity on the hazard day and temperature on the day before the hazard day as well as day of the week. 95% CL = 95% confidence limits.

* This model contains nine pairs of symmetric bidirectional controls, one for each referent-spacing length from 6 to 14 days.

Risk Set Sampling for Case-Crossover Designs(1)

Navidi & Weinhanl(Epidemiology,2002) 13:100-105

- Develop effect estimates that are free from bias caused by time trends

1) Full stratum bidirectional design
2) Matched pair design
3) Sym. Bidirectional design
4) Semi-symmetric bidirectional design(developed)

Risk Set Sampling for Case-crossover Designs(2)

Navidi & Weinhanl(Epidemiology,2002) 13:100-105

$$P(T_k \mid R) = \frac{P(T_k \cap R)}{P(R)} = \frac{e^{\beta X_k}\pi(R \mid T_k)}{\sum_{j \in R} e^{\beta X_k}\pi(R \mid T_j)}$$

T_k :Failure time R :Risk set selected

weighted version of the standard conditional logistic regression with the quantity $\pi(R \mid T_j)$ as weights.

TABLE 1. Results of Case-Crossover and Quasi-Likelihood Analyses of Simulated Data

Method	Without Seasonal Confounding		With Seasonal Confounding	
	Mean log RR	Standard Deviation	Mean log RR	Standard Deviation
FSBI	0.0990	0.0400	0.2072	0.0371
RMP	0.1006	0.0555	0.2106	0.0525
SBI	0.1645	0.0571	0.1686	0.0545
SSBI	0.1004	0.0551	0.1011	0.0517
QL	0.1017	0.0397	0.1945	0.0360

The true value of log RR is 0.1. FSBI = full-stratum bidirectional case-crossover design, in which all nonfailure times were used as controls. RMP = random matched-pair design, in which a single nonfailure time was chosen at random to be the control. SBI = symmetric bidirectional design, in which two control times were selected, both 1 week before and 1 week after failure. SSBI = semisymmetric bidirectional design, in which a single control time was selected, either 1 week before or 1 week after failure. QL = a quasi-likelihood extension of Poisson regression in which a parameter for overdispersion was included, the residuals were assumed to follow a Markov process, and mortality counts lagged 1 and 2 days were included as covariates. Results are based on 1,000 iterations.

Increased Particulate Air Pollution and the Triggering of Myocardial Infarction

Peters, Dockery, Muller, Murray, Mittleman(Circulation, 2001) 103: 2810-2815

- Myocardial Infarction onset (772 patients)

  OR 1.48 associated with an increase of 25㎍/㎥ $PM_{2.5}$ during a 2-hour period before the onset,

  & an OR of 1.69 for an increase of 20 ㎍/㎥ $PM_{2.5}$ in the 24-hour period 1 day before the onset

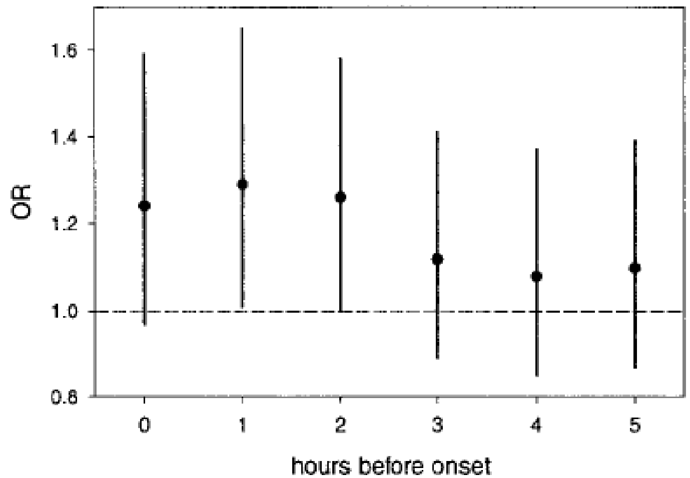


Figure 1. Univariate analyses for association between onset of MI and hourly concentrations of $PM_{2.5}$. Odds ratios and 95% CIs for an increase of 25 $\mu g/m^3$ $PM_{2.5}$.

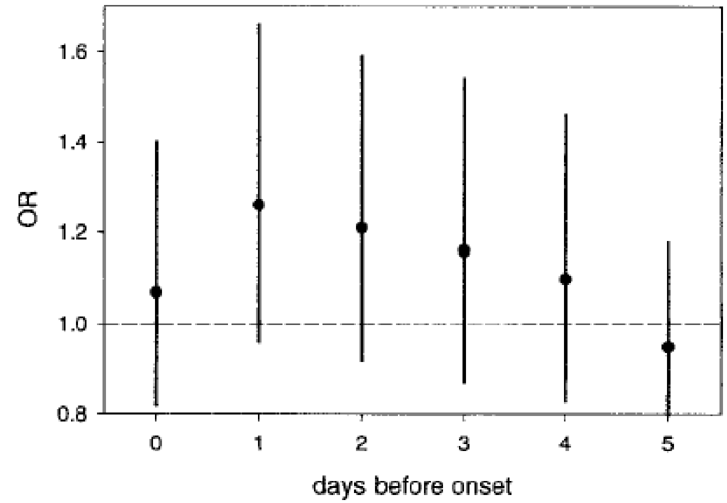


Figure 2. Univariate analyses for association between onset of MI and 24-hour average concentrations of $PM_{2.5}$. Odds ratios and 95% CIs for an increase of 20 $\mu g/m^3$ $PM_{2.5}$.

TABLE 4. Odds Ratios for 2-Hour and 24-Hour Average Concentrations of Single Pollutants Estimated Jointly.

	Increase (5th to 95th Percentile)	Unadjusted OR (95% CI) (n=772)	Adjusted* OR (95% CI) (n=764)
Particles			
$PM_{2.5}$, μg/m^3			
2-hour	25	1.43 (1.13, 1.81)	1.48 (1.09, 2.02)
24-hour	20	1.44 (1.11, 1.86)	1.62 (1.13, 2.34)
PM_{10}, μg/m^3			
2-hour	40	1.45 (1.11, 1.88)	1.51 (1.06, 2.15)
24-hour	30	1.31 (0.99, 1.73)	1.66 (1.11, 2.49)
Coarse mass, μg/m^3			
2-hour	15	1.13 (0.92, 1.40)	1.16 (0.89, 1.51)
24-hour	15	1.18 (0.85, 1.64)	1.39 (0.89, 2.15)
Black carbon, μg/m^3			
2-hour	3	1.32 (1.06, 1.65)	1.27 (0.97, 1.68)
24-hour	2	1.08 (0.84, 1.39)	1.21 (0.87, 1.70)
Gases			
Ozone, ppb			
2-hour	45	1.05 (0.76, 1.46)	1.31 (0.85, 2.03)
24-hour	30	1.21 (0.88, 1.67)	0.94 (0.60, 1.49)
Carbon monoxide, ppm			
2-hour	1.0	1.27 (0.98, 1.63)	1.22 (0.89, 1.67)
24-hour	0.6	0.99 (0.77, 1.27)	0.98 (0.70, 1.36)
Nitrogen dioxide, ppm			
2-hour	0.040	1.20 (0.91, 1.59)	1.08 (0.76, 1.53)
24-hour	0.030	1.03 (0.77, 1.39)	1.19 (0.81, 1.77)
Sulfur dioxide, ppm			
2-hour	0.020	1.00 (0.87, 1.14)	0.96 (0.83, 1.12)
24-hour	0.020	0.92 (0.71, 1.20)	0.91 (0.67, 1.23)

Estimates are calculated for a change from 5th to 95th percentile of the pollutants.
*Adjusted for season, meteorological parameters, and day of the week.

Case-crossover design
Matched case-control design
Conditional logistic regression

$\text{data}(Y_{k_j}, X_{k_j}, Z_k)$ $j = 1, \cdots, n_k$: # of observation
$k = 1, \cdots, k$: strata

Y_{k_j} : binary outcome
X_{k_j} : covariates
Z_k : stratum index

$$g(X_{k_j}, Z_k) = \beta_0 + \alpha_k + \beta' X_{k_j}$$

different intercept α_k
same slope β

k

Conditional Likelihood for the k[th] Stratum: Prob observed data conditional on the stratum total sample size and the total # of cases

- Contribution to the conditional likelihood for the k– stratum

$$\ell_k(\beta) = \frac{\prod_{\text{all cases}} \Pr(y_{k_i} = 1 \mid x) \prod_{\text{all controls}} \Pr(y_{k_i} = 0 \mid x)}{\sum_j \left[\prod_{\text{all cases}} \Pr(y_{k_i} = 1 \mid x) \prod_{\text{all controls}} \Pr(y_{k_i} = 0 \mid x) \right]}$$

- The full conditional likelihood

$$\ell(\beta) = \prod_{k=1}^{K} \ell_k(\beta)$$

In the conditional logistic regression

$$\ell_k(\beta) = \frac{\prod_{\text{all cases}} \exp(\beta' X_{k_j})}{\sum_j \left[\prod_{\text{all cases}} \exp(\beta' X_{k_j}) \right]}$$

- One-to-one match 의 경우

$$\ell_k(\beta) = \frac{\exp(\beta' X_{k_1})}{\exp(\beta' X_{k_1}) + \exp(\beta' X_{k_0})} = \frac{1}{1 + \exp\left\{\beta'(X_{k_0} - X_{k_1})\right\}}$$

Proportional hazard model 과 동일

→proc phreg으로 풀 수 있다

- Proc PHREG is to fir proportional hazard model for survival analysis
- It uses hazard function and partial likelihood

$$h_i(t) = \lambda_0(t)\exp(\beta_1 X_{i1} + \cdots + \beta_k X_{ik})$$

$$PL = \prod_{\text{All i}} \frac{\exp(\beta' X_i)}{\sum_{\text{j in Risk Set}} \exp(\beta' X_j)}$$

Example) proc PHREG

Obs	date	ID	dumtime	case	tem	hum	ap
1	05JAN98	000253306A	1	1	2.2500	74.750	10199.00
2	05JAN98	000253306A	2	0	.	.	.
3	05JAN98	000253306A	2	0	3.2125	84.250	10230.13
4	06JAN98	000171215A	1	1	-1.3750	73.250	10206.88
5	06JAN98	000171215A	2	0	.	.	.
6	06JAN98	000171215A	2	0	2.0250	84.125	10217.63
7	07JAN98	000253914A	1	1	-0.5125	58.000	10220.38
8	07JAN98	000253914A	2	0	.	.	.
9	07JAN98	000253914A	2	0	0.5375	88.875	10258.00
10	07JAN98	000253926A	1	1	-0.9250	58.125	10232.88
11	07JAN98	000253926A	2	0	.	.	.
12	07JAN98	000253926A	2	0	0.7625	87.625	10262.50
13	18JAN98	000104569A	1	1	1.0000	61.500	10227.50
14	18JAN98	000104569A	2	0	2.5750	78.625	10247.88
15	18JAN98	000104569A	2	0	-8.4125	47.625	10232.75

```
proc phreg data=comm nosummary;
model dumtime*case(0) = tem  hum ap
/ties=discrete ;


strata id;


run;
```

On Going Study 1

Stroke vs. air pollution

Triggering of Ischemic Stroke Onset by Decreased Temperature by Yun-Chul Hong et al.

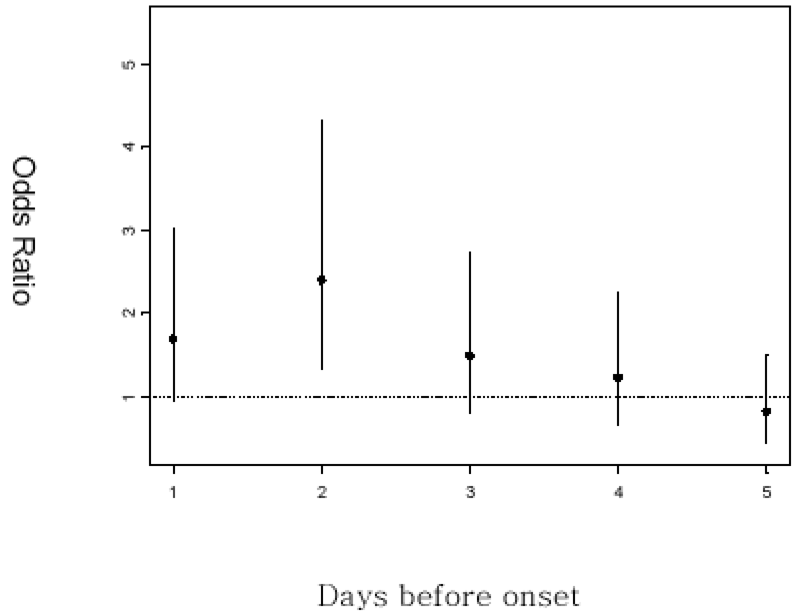
- No associations between stroke & air pollution were found
  ⇒ stroke and weather
- 1 case period is matched with 2 controls exactly 1 week apart before and after the date and time of the onset of the ischemic stroke
- 545 patients Jan 1998 – Dec 2000

On Going Study 1

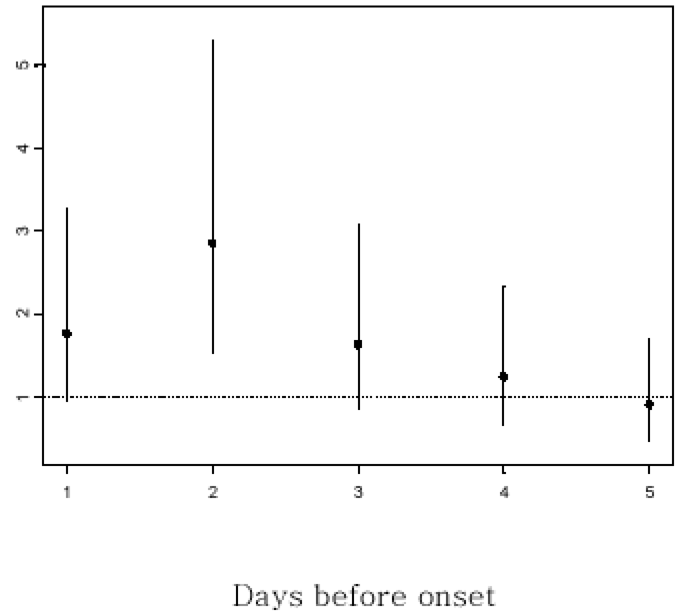
Stroke vs. air pollution

- OR=2.38 (1.33-4.34) for IQR (17.4 C) decrease of temperature
- Elevated risk period = 24 to 54 hours after the exposure to cold
- Greater in winter
- Women, elderly, pt with hypertension, hypercholesterolemia, no prior history of stroke are more susceptible

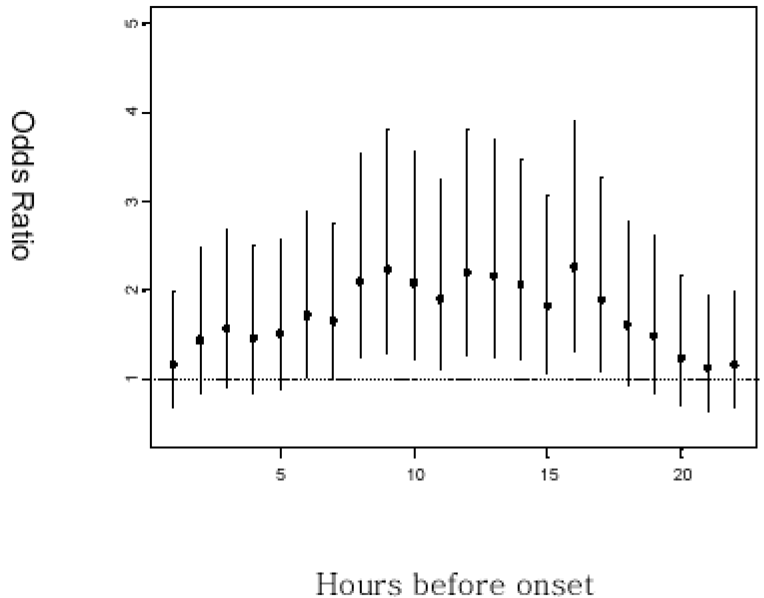
Risk factors	Adjusted model	Unadjusted model
Age <=65	2.34 (0.04)	2.09 (0.06)
Age > 65	4.03 (<0.01)	2.97 (0.02)
Male	2.39 (0.03)	1.81 (0.12)
Female	3.74 (<0.01)	3.73 (<0.01)
No prior stroke history	3.10 (<0.01)	2.66 (<0.01)
Prior stroke history	2.05 (0.03)	1.59 (0.49)
No hypertension	2.66 (0.10)	1.65 (0.32)
Hypertension	3.33 (<0.01)	3.03 (<0.01)
No hypercholesterolemia	2.39 (0.01)	2.02 (0.03)
Hypercholesterolemia	9.35 (0.02)	8.08 (0.02)
No Obesity	3.65 (<0.01)	2.48 (0.01)
Obesity	1.83 (0.31)	2.21 (0.17)
Non-Smoker	4.80 (<0.01)	4.74 (<0.01)
Ex-Smoker	4.35 (0.13)	2.56 (0.30)
Current Smoker	1.34 (0.55)	1.08 (0.88)



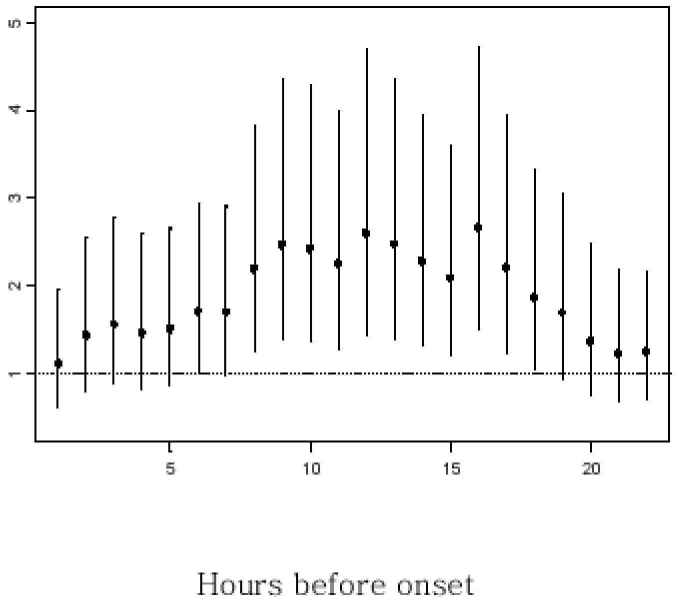
Unadjusted model



Adjusted model for humidity and air pressure



Unadjusted model



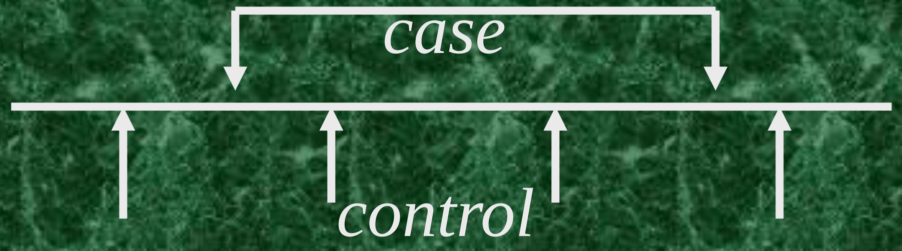
Adjusted model for humidity and air pressure

On Going Study 2

Asthma vs. air pollution

- Challenge: some patients has multiple outcomes
- Approaches

1) Ignore multiple outcome (use first outcome only)
2) Ignore subject effect
   (treat 2nd outcome as different patients)
3) Use m:2m matching rather than 1:2 matching



4) m:2m matching with subjects give some structure for controls and cases

→ numerator of l_i is different likelihood

→ Standard software doesn't work

5) Applying GEE with PHREG

Use conditional likelihood approach

- Conditional likelihood (m:2m) with single case

Let $X_k, X_{k-}, X_{k+},$: case, −control, +control

$$l_i = \frac{\exp(\beta X_k)}{\exp(\beta X_{k-}) + \exp(\beta X_k) + \exp(\beta X_{k+})}$$

$$= \frac{1}{1 + \exp(\beta(X_{k-} - X_k)) + \exp(\beta(X_{k+} - X_k))}$$

- Conditional likelihood (m:2m) with double cases

Let X_{k1}, X_{k1-}, X_{k1+},: case1, −control1, +control1

X_{k2}, X_{k2-}, X_{k2+}, case2, −control2, +control2

$$l_i = \frac{\exp(\beta X_{k1})\exp(\beta X_{k1})}{Den}$$

Usual Phreg, Den = ${}_6C_2$ terms i.e.

$$Den = \exp(\beta X_{k1-})\exp(\beta X_{k1}) + \cdots + \exp(\beta X_{k2})\exp(\beta X_{k2+})$$

In our case, Num = ${}_3C_2$ + ${}_3C_2$ terms i.e.

$$
\begin{aligned}
Num = {} & \exp(\beta X_{k1-})\exp(\beta X_{k1}) + \exp(\beta X_{k2-})\exp(\beta X_{k2}) \\
& + \exp(\beta X_{k1-})\exp(\beta X_{k1+}) + \exp(\beta X_{k2-})\exp(\beta X_{k2+}) \\
& + \exp(\beta X_{k1})\exp(\beta X_{k1+}) + \exp(\beta X_{k2})\exp(\beta X_{k2+})
\end{aligned}
$$

In general, for M cases per patients

M ${}_3C_2$ terms needed in the denominator rather than ${}_{3M}C_2$

Newton-Raphson algorithm for estimating regression Parameters

Score function (gradient) $U(\beta) = \frac{\partial l}{\partial \beta}$

Information matrix (Hessian) $I(\beta) = \frac{\partial^2 l}{\partial \beta \partial \beta'}$

$$\beta_{j+1} = \beta_j - I^{-1}(\beta_j) U(\beta_j)$$

Repeat until no change

Scheme of simulation study

- Generate correlated Binary time series outcomes
- Apply Naïve and new methods
- Compare the results

 – Bias and variance (Mean Squared Error)

Actual Problems

- Not significant association between air pollution and asthma -> increase # patients (practical ?)
- Humidity was found to be very significant ($p<0.01$) in the preliminary analyses -> focus on humidity
- Any idea, Please !!

SUMMARY
Case-Crossover Analysis

- Convenient tool
- Some problems reported
- Generally accepted methodology in environmental studies if properly done
- Simulation studies needed
- Extension to various field is possible

THANK YOU

Email : hokim@snu.ac.kr

This file is available at

http://plaza.snu.ac.kr/~hokim

열린 강의실, 세미나자료