대한지역사회영양학회지 6(5) : 945~949, 2001
Korean J Community Nutrition 6(5) : 945~949, 2001

□ 특별기획 □

영양연구 모형에 적합한 통계기법 : 영양역학에서 방법론적인 관심사

남 정 모

연세대학교 의과대학 예방의학교실

Methodological Issues for Nutritional Epidemiology

Chung-Mo Nam

Deparment of Preventive Medicine and Public Health, Yonsei University College of Medicine, Seoul, Korea

최근들어 영양역학(nutritional epidemiology)의 중요성이 강조되고 있다. 영양역학이란 식이(diat)와 질병발생과의 인과관계를 밝히는 분야로 정의할 수 있다. 이러한 영양역학에서 방법론적으로 많이 연구되는 부분은 식이섭취로부터 영양소를 측정할 때 동반되는 측정오차(measurement error)와 특정 영양소와 질병발생의 통계적 모형에서 어떤 방법으로 총열량의 효과를 통제할 수 있는가에 대한 문제이다. 본 논문은 위의 두 가지 내용의 방법론적인 측면에만 초점을 두고자 한다.

측정오차

1. 측정오차의 정의 및 분류

영양역학연구에서는 개인의 영양소 섭취를 측정하기 위해 식품섭취빈도조사지(food frequency questionnaires : FFQ)를 가장 많이 사용하고 있다. 이러한 이유는 만성질병의 발생이 어느 한 시점의 식이섭취 상태와 관련있기보다는 질병발생 이전의 식이섭취와 관련이 높기 때문이다. 많은 연구에서 FFQ를 통해 측정한 영양소 섭취는 상당한 크기의 측정오차가 있다고 보고하고 있다.

최근까지 영양역학에 대한 많은 연구들이 발표되고 있지만, 통계학적으로 유의한 영양소섭취와 질병발생과의 관계를 보여주는 연구는 그리 많지가 않다. 실제로 이들간의 관계가 유의함에도 불구하고 이를 통계학적으로 입증하지 못하는 이유들 중 식이섭취를 측정할 때 발생하는 측정오차를 대표적으로 생각할 수 있다.

측정오차는 크게 랜덤(random)과 계통적(systemic) 오차로 나눌 수 있다. 측정값의 평균이 참값이 되는 경우를 랜덤오차로 그렇지 않으면 계통적 오차로 정의한다. 이러한 오차의 분산은 개인내(within person) 분산과 개인간(between person) 분산의 성분으로 나누어 설명할 수 있다. 개인내 분산이란 각 개인의 참값으로 부터 발생하는 개인내의 자료의 변동을 의미하며, 예를들어 개개인의 영양소 섭취가 일별로 차이가 있으므로 발생한다고 볼 수 있다. 개인간 분산이란 집단내에서 각 개인의 참 값들의 차이에서 발생하는 변동으로 정의할 수 있다.

2. 측정오차로 인한 효과

측정오차로 인한 영향을 생각할 때 그 변수가 연속형이냐 범주형이냐에 따라 사용되는 용어 및 그 효과가 다르다. 따라서 영양소를 연속형으로 분석하는 경우와 범주화하여 분석하는 경우로 나누고, 측정오차가 랜덤인 경우에만 제한하여 설명한다.

(1) 연속형으로 분석하는 경우

일반적으로 연속형으로 측정된 영양소 섭취가 랜덤 측정오차를 동반한다면 영양소섭취와 질병발생의 관계에 대한 추정치는 치우침(bias)을 가지며 그 치우침의 방향은 "null"을 향한다(towards the null)고 알려져 있다. 이에 대한 내용을 먼저 다음과 같이 정의한 측정오차모형을 이용하여 설명하도록 한다.

$$z = x + \varepsilon$$

여기서, 1회 측정한 특정영양소의 관측값 z에 대해, x는 참값으로서 평균이 μ, 분산이 σ_b^2(개인간 분산)인 분포를 따르고, ω은 측정오차로서 평균이 0, 분산이 σ_w^2(개인내 분산)인 정규분포를 따르고 측정오차와 x는 서로 독립임을 가정한다. 만약 연속형으로 측정한 종속변수를 y라 할 때, 특정영양소의 관측값 z와 y의 관찰된 상관계수(r_{obs})는 참 상관계수(r_{true})와 다음과 같은 관계가 있다.

$$r_{true} = r_{obs}\sqrt{(1+\sigma_w^2/\sigma_b^2)} \quad \text{(식 1)}$$

즉, 관찰된 상관계수는 참상관계수보다 항상 작으며, 그 정도는 개인내 분산과 개인간 분산의 비에 따라 변함을 알 수 있다. 만약, 개인내 분산이 0이면 참 상관계수와 관찰된 상관계수는 같음을 알 수 있고, 특정영양소를 m번 측정하여 평균을 사용하면 개인내 분산이 σ_w^2/m 크기로 작아지므로 치우침의 정도를 줄일 수 있다. Liu 등(1977)은 상관계수의 정확도가 어느 정도 되기 위해서는 몇번을 반복적으로 측정하여야 하는지를 보고하였다.

한편 특정 영양소들의 개인내 분산과 개인간 분산에 대하여 그동안 많은 연구들이 있었으며, Wilkens & Lee는 이들 연구결과들로부터 특정 영양소들의 개인내 분산과 개인간 분산의 비를 종합하였다(Table 1).

즉, Table 1로부터 식이섭취를 통한 영양소 섭취는 개인간 분산보다 개인내 분산이 더 크며, 특히 Vitamin A인 경우 총분산 중 개인내 분산이 82.1% 정도로 매우 높음을 알 수 있다. 위의 분산의 비를 고려할 때, 특정영양소와 질병발생의 관련성을 연구하는 영양역학에서 개인내분산을 고려하는 것이 얼마나 중요한지 알 수 있다.

측정오차의 영향을 줄이기 위한 방법으로 첫번째 생각할 수 있는 것은 식이섭취를 여러번 측정하여 이들의 평균값을 사용하는 방법이다. 다음으로 "calibration study"를 통하여 통계적인 방법으로 측정오차의 영향을 줄일 수 있으나 식이섭취를 통한 특정영양소의 참 값을 측정할 수 있어야 하는 어려운 점이 있다. 기타 많은 통계적인 방법들이 계속적으로 제안되고 있으나 이에 대한 내용은 생략하기로 한다.

(2) 범주화하여 분석하는 경우

기본적으로 식이를 통한 영양소 섭취는 연속형 변수로 측정된다. 그러나 예전부터 많은 연구에서 통계적 효율(efficiency)의 손실에도 불구하고 특정영양소를 다음과 같은 이유 등으로 범주화하여 분석하고 있다. 범주화하여 분석하면 그 결과를 해석하는 것이 직관적이고, 또한 그 결과를 인구집단에 적용하는데 용이하다. 그리고 특정 영양소와 질병발생의 통계적 모형을 쉽게 설정할 수 있고, 영양소 섭취가 아주 작거나 큰 이상점의 영향에 덜 민감하기 때문이다. 지금부터의 논의는 식이를 통한 특정 영양소를 사분위(quartile) 또는 권장량을 기준으로 이분(binary)변수 등과 같이 범주화 하였다고 가정한다.

Table 1. Ratio of within person to between person variance

Nutrient	No. of studies	Median ration	Within person variance of the total variance(%)
Energy (Kcal)	12	1.4	58.3
Protein	10	1.4	58.3
Carbohydrate	9	1.2	54.5
Fat	9	1.3	56.5
% of calories from fat	8	2.4	70.6
Saturate fat	8	1.5	60.0
Cholesterol	11	4.4	81.5
Vitamin C	9	2.3	69.7
Vitamin A	7	4.6	82.1
Iron	8	2.4	70.6
Calcium	10	1.6	61.5
Zinc	6	2.4	70.6
Dietary fiber	3	1.7	63.0

범주형자료에서 측정오차로 인해 범주가 잘못 분류되는 것을 분류오류(misclassification)가 발생하였다고 하며, 무차별(nondifferential)과 차별(differential) 분류오류로 나눌수 있다. 예를 들어 환자-대조군 연구에서 환자군과 대조군의 영양소 섭취의 범주에 대한 분류오류가 동일하면 무차별 분류오류, 그렇지 않으면 차별 분류오류로 정의한다. 일반적으로 이분형인 경우 무차별 분류오류가 발생하면 영양소와 질병발생의 관계에 대한 추정치의 치우침은 "null"을 향한다. 그러나 차별분류오류인 경우는 추정치의 치우침에 대한 방향이 "null"을 향할 수도 있고, "null"로부터 멀어지는(away from the null) 방향으로 발생할 수 있다. 한편 영양소를 사분위수로 범주화하는 경우와 같이, 세 개 이상의 범주를 갖는 경우에는 분류오류의 형태에 관계없이 영양소와 질병발생에 대한 추정치의 치우침은 그 방향이 일정하지 않다고 알려져 있다.

분류오류를 구분하는 이유는 만약 분류오류가 무차별 분류오류라고 확신할 수 있고, 영양소섭취가 이분형으로 범주화되었다면 영양소와 질병발생의 관련성의 크기는 과소추정되었고 따라서 이와 같은 상황에서 그 관련성이 통계학적으로 유의하였다면 실제로 그 관계는 유의하다고 판단할 수 있기 때문이다. 그러나 위와 같은 이분형인 경우로 범주화한 경우에도 방법론적인 문제가 있다. 즉, 연속형으로 측정된 영양소는 앞서 언급한 대로 측정오차를 동반하며 이를 이분화하는 경우에 분류오류가 발생하고 그 형태는 무차별 분류오류 뿐 아니라 차별 분류오류 형태로 발생할 수 있다.

분류오류로 인한 추정치의 치우침을 줄이고자 그동안 많은 방법론적인 연구가 진행되었으며 이에 대한 논의는 생략하기로 한다.

총열량을 통제하는 방법

1. 총열량을 통제하여야 하는 이유

식이와 만성질환의 관련성에 관한 관심이 증대됨에 따라

식이 섭취를 측정하는 조사도구의 개발이 활발해졌으나 이에 비하여 식이 자료를 분석하고 해석하는 방법에 대한 연구는 상대적으로 소홀했으며, 총열량의 의미는 종종 간과되어 왔다. 그 결과 특정 영양소가 질병의 발생에 미치는 영향에 대한 연구결과의 해석이 난해해졌다. 이처럼 영양역학 연구에서는 연구자가 매우 신중하게 식이 자료를 수집하였다 하더라도 자료를 분석하고 해석하는 중에 잘못된 결론을 도출할 수 있는 여지가 크다. 특히 총열량은 제반 여건을 통제하는 실험연구와는 달리 체격, 활동정도 및 대사효율 등 개인적인 차이에 영향을 받고 있는 바, 이를 고려하지 않고 질병과 영양소의 관계를 규명할 경우 그 결과는 실제와 상반되게 나타날 수 있다. 또한 대상자가 섭취한 식품의 영양소 함량에 따라 대상자의 영양소 섭취량이 계산되어지는데, 대부분의 영양소 섭취는 열량섭취와 관련이 있으며, 영양소 섭취량 평가시의 오차는 총열량섭취의 오차와 상관되어 특정 영양소의 섭취량 변동에 영향을 준다. 따라서 특정 영양소와 질병의 관련성을 밝히고자 할 때는 대상자의 총열량을 고려하여야 한다.

영양역학 연구에서 총열량을 통제해야 하는 근거와 잇점은 다음과 같다.

1) 혼란 효과를 통제

특정 영양소와 질병과의 관련성을 규명하고자 할 때 만약 총열량이 질병의 직접적인 결정요인으로 작용한다면 즉, 체격, 활동정도 및 대사효율 등이 질병 발생과 연관되어 있는 경우, 이때 총열량은 혼란변수로 작용하게 된다. 따라서 총열량의 영향을 통제하지 않으면, 특정 영양소와 질병의 관련성이 왜곡될 수 있다.

2) 외적변동(Extraneous variation)을 제거

특정 영양소와 질병과의 관련성을 규명하고자 할 때 총열량이 질병의 직접적인 원인은 아니지만 질병 발생과 관련있을 수도 있다. 이는 곧 체격, 신체활동정도 및 대사효율 등이 질병 발생의 주된 결정요인이 아닌 경우를 의미하는데, 이러한 요인들은 질병과 직접적인 관련이 없다 하더라도 영양소 섭취의 변동에 기여하기 때문에 특정 영양소와 질병의 관계 해석에 외적인 변동을 주게 된다. 따라서 이와 같은 경우 체격, 활동정도 및 대사효율 등을 측정하여 이의 영향을 통제한 후 특정 영양소와 질병의 관련성을 규명하고자 하는 시도가 제시된 바 있다. 그러나 이러한 생물학적인 변수들은 대규모 인구집단을 대상으로 하는 역학연구에서 불가능하다. 이에 생물학적인 변수에 의해 영향을 받는 총열량을 보정하여 특정 영양소와 질병의 관계에 대한 생물학적 요인의 영향을 간접적으로 통제한다.

2. 총열량의 효과를 통제하는 회귀모형 및 방법론적 문제점

특정 영양소(F)와 질병(D)의 관련성을 알아보기 위해 총열량(T)를 통제하는 회귀분석 방법은 다음과 같은 4가지 모형을 많이 사용하고 있다. 먼저, 질병 D에 대한 연결함수(link function)를 M(D)라 정의하자. 해석의 편의상 F을 지방(fat)으로 인한 열량, 그리고 M(D)를 로지스틱 회귀모형에서의 로짓 연결함수, 즉 $M(D)=\log(P/(1-P))$로 가정하자. 여기서 P는 특정 독립변수에서 질병에 걸릴 확률이다.

1) 표준모형(Standard model)

$$M(D)=\beta_{0S}+\beta_{1S}F+\beta_{2S}T+\omega$$

이 모형은 회귀모형에 총열량을 하나의 독립변수로 포함시킨 모형이다. 관심 있는 회귀계수 β_{1S}는 총열량이 일정할 때 지방섭취가 1(kcal) 증가하면, 질병에 걸릴 위험(RR)은 $\exp(\beta_{1S})$은 증가한다는 의미이다. 여기서, 총열량이 일정할 때 지방섭취가 1 증가한다는 것은 결국 단백질이나 탄수화물과 같은 지방이외의 열량원(non-fat)섭취가 1 감소하여야 하므로 이 효과는 비지방 1 kcal가 지방 1 kcal로 대체(substitute)된 효과로 볼 수 있다. 즉, 순수한 지방섭취만의 효과가 아니다.

2) 잔차모형(Residual model 또는 willett method)

$$M(D)=\beta_{0R}+\beta_{1R}R+\beta_{2R}T+\omega$$

여기서, R은 F을 종속변수로, T를 독립변수로 한 회귀모형에서 추정된 잔차이다. 즉, 지방섭취에서 총열량의 차이로 인한 영향을 제거한 후의 크기로 생각할 수 있다. 따라서 이 모형으로 자료를 분석하기 위해서는 먼저 R을 추정하고 다음 단계로 R과 T를 독립변수로 하여 회귀모형을 구축하면 된다. 앞서 설명한 것과 동일하게 회귀계수 β_{1R}의 의미도 대체효과로 볼 수 있다. 또한 잔차모형에서는 열량 보정시의 영양소 섭취량 산출도 가능한데, 잔차에 어떤 상수값을 (논란이 있기는 하지만 대개의 경우 평균 섭취을 사용) 더하여 개인의 영양소 섭취량을 가늠하게 된다.

3) 분할모형(Partition model)

$$M(D)=\beta_{0P}+\beta_{1P}F+\beta_{2P}(T-F)+\omega$$

분할모형은 지방과 지방이외의 열량원으로 인한 열량을 회귀모형의 독립변수에 추가한 모형이다. 여기서, T-F은 지방이외의 열량원에 의한 열량이다. 분할모형에서의 회귀계수 β_{1P}는 지방이외의 열량원이 일정할 때 지방이 1 kcal 증가하면 질병에 걸릴 위험이 $\exp(\beta_{1P})$ 만큼 증가한다는 지방의 추가(add) 섭취효과로 볼 수 있다. 그러나 이 경우 해석상에서 총열량 섭취도 증가된다는 문제점과 관심 있는 영양소가 vitamin C와 같이 micronutrient인 경우는 적용

할 수가 없다.

4) 영양밀도모형(Nutrition density model)

$M(D)=\beta_{0N}+\beta_{1N}(F/T)+\beta_{2N}T+\omega$

밀도모형은 총열량 중 지방이 차지하는 분율과 총열량을 독립변수로 한 회귀모형이다. 회귀계수 β_{1N}의 해석은 총열량의 역수에 대한 효과까지도 포함하고 있음을 알 수 있다. 이 모형은 만약 총열량이 질병발생과 관계없다면 독립변수인 지방이 차지하는 분율은 앞서 설명한 체격, 활동정도, 그리고 대사의 개인간 차이에 대한 변동을 반영하므로 바람직하다고 볼 수 있다. 또한 식이 지침 설정에 용이한 자료를 제공해 주지만, 총열량이 질병발생과 관계 있는 경우에는 해석의 어려운 점이 있다.

한편 영양밀도모형을 제외한 나머지 모형에서 추정되는 특정영양소의 회귀계수는 대체효과 또는 추가섭취효과로서 관심 있는 특정영양소의 단독효과로 설명할 수 없음을 지적하였다. Wacholder 등(1994)은 다음 모형을 사용하여 위의 3가지 모형의 F에 해당하는 회귀계수의 크기를 설명하고자 하였다.

$M(D)=\beta_0+\beta_F F+\beta_{(T-F)}(T-F)+\beta_T T+\omega$

(위의 모형은 완전한 다중공선성(multicollinearity)이 존재하기 때문에 회귀계수를 추정하는 것이 불가능하다). 이 모형을 표준모형, 분할모형, 그리고 잔차모형의 형태으로 변환하면 다음과 같은 회귀계수의 관계식이 주어진다.

$\beta_{1S}=\beta_F-\beta_{T-F}$, $\beta_{1P}=\beta_F+\beta_T$, $\beta_{1R}=\beta_F-\beta_{T-F}$

즉, 표준모형, 분할모형, 그리고 잔차모형에서 추정된 지방섭취의 효과는 지방이외의 열량원 또는 총열량의 효과가 같이 포함되어 있음을 알 수 있다. 영양역학에서 이 부분이 아직까지도 남아있는 방법론적인 큰 문제점이며 추후 계속적인 연구가 필요한 부분이라 할 수 있다.

참 고 문 헌

남정모 · 오희철(1993) : 환자 - 대조군 연구에서 무차별 분류오류에 대한 연구. *한국역학회지* 15(1) : 85-95

남정모 · 강형곤 · 서 일(1996) : 측정오차를 동반한 폭로변수와 분류오류에 대한 연구. *한국역학회지* 18(1) : 108-118

남정모 · 이선희 · 박형욱(1996) : 생태학적 자료에서 분류오류의 영향과 모형선택. *한국역학회지* 18(2) : 142-150

Beaton GH, Milner J, Corey P, McGuire V, Cousins M, Stewart E, de Ramos M, Hewitt D, Grambsch PV, Kassim N, Little JA (1979) : Sources of variance in 24-hour dietary recall data : implications for nutrition study design and interpretation. *Am J Clin Nutr* 32(12) : 2546-2549

Birkett(1992) : Effect of nondifferential misclassfication on estimates of odds ratios with multiple levels of exposure. *Am J Epidemiol* 136(3) : 356-362

Brown CC, Kipnis V, Freedman LS, Hartman AM, Schatzkin A, Wacholder S(1994) : Energy adjustment methods for nutritional epidemiology. *Am J Epidemiol* 139(3) : 323-338

Copeland KT, Cheskoway H, Mcmichael AJ, Holbrook RH(1977) : Bias due to misclassfication in the estimation of relative risk. *Am J Epidemiol* 105(5) : 488-495

Dosemeci M, Wacholdet S, Lubin JH(1990) : A brief original contribution : Does nondifferential misclassfication of exposure always bias a true effect toward the null value?. *AM J Epidemiol* 132(4) : 746-748

Freudenheim JL(1999) : Study design and hypothesis testing : Issues in the evaluation of evidence from research in nutritional epidemiology. *Am J Clin Nutr* 69 : 1315-1321

Greenland S(1980) : The effect of misclassfication in the presence of covariates. *Am J Epidemiol 112(4) : 564-569*

Howe GR, Miller AB, Jain M(1986) : Re : Total energy intake : Implication for epidemiologic analyses. *Am J Epidemiol* 124 : 157-159

Hu FB, Stempfer M, Rimm E, Ascherio A, Rosner B, Spiegeiman D, Willett W(1999) : Dietary fat and coronary heart disease : A comparison of approaches for adjusting for total energy intake and modeling related dietary measurements. *Am J Epidemiol* 149(6) : 531-540

Hunter DJ, Spiegelman D, Adami HO, Beeson L, van den Brandt PA, Folsom AR, Fraser GE, Goldbohm RA, Graham S, Howe GR, Kushi LH, Marshall JR, McDermott A, Miller AB, Speizer FE, Wolk A, Yaun SS, Willett W(1996) : Cohort Studies of Fat Intake and the Risk of Breast Cancer? A Pooled Analysis. *N Engl J Med* 334(6) : 356-361

Kipnis V, Freedman LS, Brown CC, Hartman A, Schatzkin A, Wacholder(1993) : Interpretation of energy adjustment models for nutritional epidemiology. *Am J Epidemiol* 137(12) : 1376-1380

Kupper LL(1984) : Effects of the use of unreliable surrogate variables on the validity of epidemiologic research studies. *Am J Epidemiol* 120(4) : 643-648

Liu K, Stamler J, Dyer A, Mckeever J, Mckeever P(1978) : Statistical methods to assess and minimize the role of intra-individual variability in obscuring lipids and serum cholesterol. *J Chron Dis* 31(6-7) : 399-418

Marshall JR, Priore R, Graham S, Brasure J(1981) : On the distortion of risk estimates in multiple exposure level case-control studies. *Am J Epidemiol* 114(4) : 464-473

Pike MC, Bernstein L, Peters RK(1989) : Re : Total energy intake : implications for epidemiologic analyses. *Am J Epidemiol* 129(6) : 1312-1315

Pike MC, Peters RK, Bernstein L(1993) : Re : Total energy intake : implications for epidemiologic analyses. *Am J Epidemiol* 137(7) : 811-812

Prentice RL(1996) : Measurement error and results from analytic epidemiology : Dietary fay and breast cancer. *J Natl Cancer Inst* 88(23) : 1738-1747

Sempos CT, Liu K, Ernst N(1999) : Food and nutrient exposure : what to consider when evaluating epidemiologic evidence. *Am*

J Clin Nutr 69(6) : 1330-1338

Shekelle RB, Nichaman MZ(1987) : Re : Total energy intake implications for epidemiologic analyses. *Am J Epidemiol* 126(5) : 980-983

Wacholder S, Schatzkin A, Freedman LS, Kipnis V, Hartman A, Brown CC(1994) : Can energy adjustment separate the effects of energy from those of specific macronutrient?. *Am J Epidemiol* 140(9) : 848-855

Willett W, Stempfer MJ(1986) : Total energy intake : Implication for epidemiologic analyses. *AM J Epidemiol* 124(1) : 17-27

Willett WC, Stempfer MJ(1993) : Re : Total energy intake implications for epidemiologic analyses. *Am J Epidemiol* 137(7) : 811-813

Willett WC, Howe GR, Kushi LH(1997) : Adjustment for total energy intake in epidemiologic studies. *Am J Clin Nutr* 65S : 1220-1228

Willett W, Nutritional Epidemiology. 2nd, Oxford University Press, New York, 1998

Wilkans L, Lee J, Nutritional Epidemiology. In : Armitage P, Colton T, eds. Encyclopedia of Biostatistica, pp.3099-117, West Sussex, John Wiley & Sons