추적 관찰된 자료의 회귀분석적 방법

연세의대 남정모

○ 보건학 연구에서 추적 관찰된 자료를 분석하는 경우가 많으며, 일반적으로 추적관찰된 자료의 분석방법으로 생존분석 (survival analysis)을 많이 사용하고 있다.

○ 추적관찰된 자료의 회귀분석으로 Cox's의 비례위험회귀모형(proportional hazards model)과 Poisson 회귀모형 등을 사용할 수가 있다.

○ 한편 코호트 연구로 자료를 수집하고 비용효과적인 측면을 고려하였을 때 Nested case-control 또는 Case-cohort 연구설계로 변형하여 연구하는 경우가 있다.

○ 본 강의는 생존분석의 기본개념에 대해 고찰하고,
- Cox'의 비례위험회귀모형
- Poisson regression model
- Nested case-control의 기본개념과 이러한 연구설계에서의 회귀분석 방법
- Case-cohort의 기본개념과 이러한 연구설계에서의 회귀분석 방법을 고찰한다.

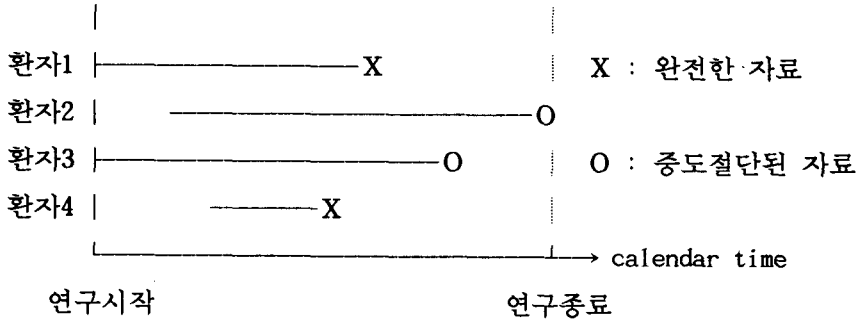
1. 생존분석 (Survival Analysis)

가. 생존분석의 특징

어떤 사건(예:사망)이 발생할 때까지의 시간(time)과 관련된 통계적 분석 방법

- 사건의 발생 여부에 대해 중도절단된 자료(censored data)가 존재
- 중도절단된 자료가 발생하는 이유

  loss to follow up, drop out, termination of the study, death from unrelated cause
- 생존분석 자료의 형태



- 생존시간의 정의에 따른 주의점
  - 생존시간의 시작점에 대한 명확한 정의를 할 수 있는가?
  - 생존시간을 정확하게 측정할 수 있는가?
  - 사건의 발생여부를 확실히 구별할 수 있는가?
- 생존분석은 중도절단된 자료의 부분적인 정보를 최대한 이용

- 중도절단된 자료가 없는 일반적인 통계적 방법과 생존분석의 비교

	일반적 방법	생존분석
자료의 요약	기술통계량 평균,중위수,분산, 범위, 히스토그램	생존함수와 위험함수의 추정 지수분포, 와이블분포 생명표방법, Kaplan-Meier 방법
k개 집단의 평균 비교	모수적 방법 t-검정, 일요인 분산분석 비모수적 방법 윌콕슨검정,크루스칼-왈리스	모수적 방법 우도비검정 비모수적 방법 로그순위검정 일반화된 윌콕슨검정
회귀분석	다중회귀분석 로지스틱회귀분석	Cox'의 비례위험회귀모형 Accelerated failure time model

나. 생존함수와 위험함수의 추정

1) 생존함수(survival function)와 위험함수(hazard function)

- t 시점에서 생존함수(S(t))는 t 시점까지 사망하지 않고 생존할 확률
- t 시점에서 위험함수(h(t))는 t 시점까지 생존한 사람이 t 시점 바로 직후 순간적으로 사망할 조건부 확률 (순간사망률)

2) 생존함수와 위험함수의 일반적 개형

- 생존함수는 시간의 경과에 따라 단조감소(monotone decreasing)
  - 5 year survival rate, median survival time
- 위험함수는 시간의 경과에 따라 일정(지수분포)
  단조감소, 단조증가(Weibull 분포), 증가하다가 감소(로그-정규분포)하는 등 일정한 pattern이 없음
- 생존함수와 위험함수는 함수관계가 있지만 반드시 (반)비례례적인 관계는 아님

3) 중도절단된 자료에서의 회귀분석적 방법

- Cox's proportional hazards model

Baseline	$h_0(t)$: 모든 독립변수가 0일 때의 위험함수
Model	t시점에서 p개의 독립변수가 $x_1, x_2, \ldots x_p$일 때의 위험함수
	$h(t,x) = h_0(t)e^{b_1x_1+b_2x_2+\cdots+b_px_p}$
Assumption	분포에 대한 가정은 없으나 비례위험에 대한 가정
Model check	독립변수의 서로 다른 값에서 $\log(-\log S(t))$와 t는 시점에 관계없이 일정함 (비례위험)
SAS 모듈	PHREG

- 비례위험 회귀모형에서 회귀계수의 해석

$$\frac{h(t,X_k=x+1)}{h(t,X_k=x)} = \frac{h_0(t)\exp(b_1X_1+\cdots+b_k(x+1)+\cdots+b_pX_p)}{h_0(t)\exp(b_1X_1+\cdots+b_k(x)+\cdots+b_pX_p)} = \exp(b_k)$$

4) 생존자료분석시 주의점

- 장기간 추적관찰된 연구인 경우, 예를들어 연구기간이 20년이라 하면 연구시작 초기에 연구에 참여한 사람과 연구후반에 참여한 사람이 경험하게 되는 환경은 여러가지로 크게 차이가 있을 것이다. 이런 경우는 연구에 참여한 calendar time으로 집단을 나누고 따로 분석하거나 calendar time으로 나눈 집단을 가변수로 처리하여 분석할 수가 있을 것이다.

- 중도절단이 발생하는 것은 치료방법이나 환자의 특성에 상관없이 독립적으로 발생하여야 한다. 만약 이 가정이 어긋나면 기존의 통계적 방법들은 편의(bias)를 가지게 된다.

- 비례위험모형 가정하에서는 생존함수 값이 크면 위험함수 값은 작아지는 관계가 있기 때문에 해석을 할 때 큰 문제가 없지만 비례위험모형의 가정이 만족하지 않거나 특히 위험함수가 어긋나는 경우(crossing hazard)는 분석시 상당한 주의를 요하게 된다.

- 시간에 따라 값이 변하는 독립변수(time dependent covariates)가 여러 개 있는 경우 앞에서 언급한 비례위험회귀모형을 그대로 사용하는 것은 부적절하며 오히려 이런 경우는 포아송 회귀분석으로 분석하는 것도 하나의 방법이 될 수가 있다.

2. Poisson regression

먼저 Poisson regression을 하기 전에 generalized linear model에 대해 간단히 살펴보자.

가. GLM(Generalized Linear Model)이란 무엇인가?

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon$$

$$\Updownarrow$$

$$E(y) = \mu = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

- Normal 분포이외에 많은 분포가 존재한다 (Exponential family).
- 로지스틱 회귀분석의 결과변수는 어떤 사건이 일어날 확률이므로 0에서 1사이에 존재하여야 한다.
- 평균에 따라 분산도 같이 변화하는 경구가 있다.

나. GLM에서의 여러 가지 component

○ Random component :
- dependent variable y 에 대한 변이
- y 의 기대치 $E(y) = \mu$

○ Systematic component :
- 주어진 p 개의 independent variable들이 선형결합식으로 표현되는 부분
- $\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$

○ Link function :
random component와 systematic component를 연결시켜주는 함수 : $g(\cdot)$

○ $var(y_i) = \phi V(\mu_i)/w_i$, ϕ : constant, w : known weight

다. GLM의 예

○ traditional linear model ("다중회귀분석") :
response var. : continuous
distribution : normal, $Var(\mu) = 1$
link function : *identity*, $g(\mu) = \mu = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$

○ logistic regression

response var. : proportion

distribution : binomial (0, 1), $Var(\mu) = \mu(1-\mu)$

link function : *logit*, $g(\mu) = logit(\mu) = \log(\frac{\mu}{1-\mu})$

$= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$

○ Poisson regression

response var. : count

distribution : Poisson, $Var(\mu) = \mu$

link function : *log*, $g(\mu) = \log(\mu) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$

라. 포아송 회귀분석의 모형 및 회귀계수의 의미

○ 특정한 "age-time-exposure" 칸에 대한 사망자 수를 $\boldsymbol{d}$ 라 할 때

$$P(\boldsymbol{d} = x) = e^{-\lambda n} \frac{(\lambda n)^x}{x!}$$

여기서, $\boldsymbol{\lambda}$ 는 미지의 rate이고 $\boldsymbol{n}$ 은 인년(person year, PY)을 나타낸다.

○ Model : $\log(\lambda_x) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$

$$\downarrow \quad \lambda_x = \frac{d_x}{PY_x}$$

$$\log(\frac{d_x}{PY_x}) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

$$\log(d_x) = \log(PY_x) + \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

○ 회귀계수 $\boldsymbol{\beta}$ 의 의미 : 위의 Model에서,

$$\begin{cases} x = x_1 = 0 : & \log(\lambda_{x1=0}) = \alpha + 0 \\ x = x_1 = 1 : & \log(\lambda_{x1=1}) = \alpha + \beta_1 \end{cases}$$

$$\log(\lambda_{x1=1}) - \log(\lambda_{x1=0}) = \log(\frac{\lambda_{x1=1}}{\lambda_{x1=0}}) = \beta_1$$

$$RR = \frac{\lambda_{x1=1}}{\lambda_{x1=0}} = e^{\beta_1}$$

즉, x_1 이 0에서 1로 증가할 때 RR는 $e^{\beta 1}$ 만큼 증가한다.

3. Nested case-control design에서의 회귀분석

가. Incidence density sampling 방법

○ 관심 있는 사건이 발생하는 각 시점에서 위험집단(at risk set: 이 시점까지 관심 있는 사건이 발생하지 않은 집단)을 구축하고, 위험집단에서 랜덤하게 대조군을 뽑는다.

○ 대규모 자료에서 control을 어떻게 뽑을 것인가?

다음 논문에서 SAS 프로그램을 이용하여 incidence density 방법으로 control을 뽑는 방법을 설명하고 있다.

"Pearce N. Incidence density matching with a simple SAS computer program. Int J Epidemiol 1989; 18: 981-4 ".

나. 분석방법

○ Case가 발생한 시점에서 "at risk"에 있는 대상을 control을 뽑았으므로 matching 되어 있는 형태이다.

○ 따라서 환자-대조군이 matching 되어있는 정보를 이용한 회귀분석을 하여야 하고 이 경우 conditional logistic을 사용할 수 있다. 짝짓기 비가 1:1이 아니고 m:n인 일반적인 경우는 discrete-time에 대한 Cox의 다음 모형을 프로그램하여 수행하는 것이 수월하다 (SAS의 PHREG procedure에서 쉽게 분석할 수 있다).

$$\log\left(\frac{p_{il}}{1-p_{il}}\right)=\beta_l+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_p x_{ip}$$

여기서, p_{il} 은 i-번째 대상이 l time 이전에 사건이 발생하지 않았다는 조건하에서 l time에서 사건이 발생할 위험함수임.

4. Case-cohort design에서의 회귀분석

가. Control을 뽑는 방법

연구시작 시점에서 코호트의 부분 코호트(sub cohort)를 구축하고, 전체 코호트의 관심있는 사건을 추적관찰 하면서 case는 전체 코호트에서 관심있는 사건이 발생한 집단, 대조군은 부분 코호트에서 사건이 발생하지 않은 집단으로 정의.

나. 분석방법

○ 일반적인 Cox의 비례위험회귀모형을 사용할 수 있다.

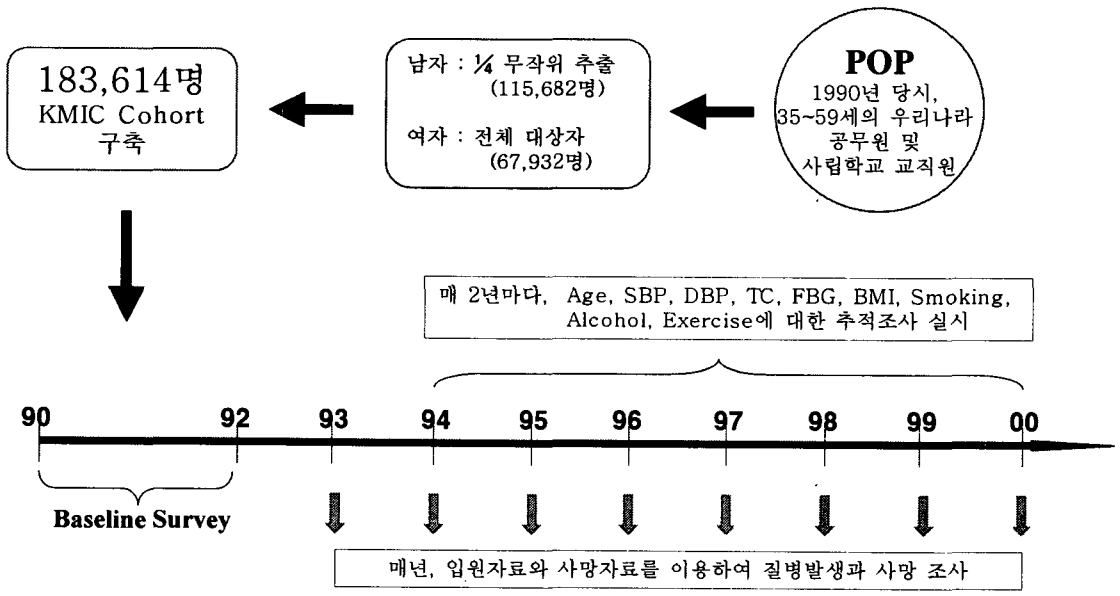
o 그러나 Cox의 비례위험회귀모형을 사용하는 경우 회귀계수의 추정에 bias가 존재할 수 있다. Bias를 제거하기 위한 방법이 "Therneau T, Li H. Computing the Cox Model for Case Cohort Design. Lifetime Data Analysis 1999; 5: 99-112"의 논문에 소개되어 있음. 특히 부분 코호트의 크기가 작고 사건이 발생하는 확률이 높으면 편의는 커질수 있다.

Case-cohort와 Nested case-control의 statistical efficiency는 비슷하다. 그러나 자료분석 과정에서 Nested case-control은 control을 뽑는 것이 어렵고 Case-cohort는 Cox의 비례위험회귀모형으로 분석할 때 분석방법을 수정하여야 하는 어려운 점이 있다. 한편 Case-cohort는 multiple outcome이 가능하다는 장점이 있다.

5. 실제 자료를 통한 분석

가. 분석에 사용된 자료

위에서 언급한 방법들을 실제 연구자료에 적용하여 그 분석방법 및 결과를 알아본다. 사용하고자 하는 자료는 KMIC(Korea Medical Insurance Corporation) Study 자료로서 연구설계는 다음과 같다.



위의 자료에서 이번 자료분석은 남자만 국한하고 또한 랜덤하게 10%만 추출하여 분석하였다. 분석자료의 변수는 다음과 같다.

변수명	설명
DCD	0=Censoring, 1=Death
SDAYS	Event가 발생할 때 까지의 Time (일)
AGE gAGE	나이(실수) 1:40세미만, 2:40~44세, 3:45~49세, 4:50~54세, 5:55세이상 ⇨ AGE2, AGE3 AGE4, AGE5
BMI gBMI	체질량지수 Body Mass Index=Weight(kg)/Height(m)^2 1:18.5미만, 2:18.5~23, 3:23이상 ⇨ BMI2, BMI3
gBP	SBP=Systolic Blood Pressure (mmHg), DBP=Diastolic Blood Pressure (mmHg) JNC7 기준에 의한,1:Normal, 2:Prehypertension, 3:Hypertension ⇨ BP2, BP3
gSMOK	1: Non-smoker, 2: Ex-smoker, 3: Current smoker ⇨ SMOK2, SMOK3

나. 분석프로그램 및 결과

1) Cox의 비례위험회귀모형

```
PROC PHREG DATA=KDR.A10;
   MODEL SDAYS*DCD(0)=AGE2 AGE3 AGE4 AGE5 BP2 BP3 EXSMOK CUSMOK BMI2 BMI3 / RL;
RUN;
```

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits	
AGE2	1	0.56691	0.21120	7.2050	0.0073	1.763	1.165	2.667
AGE3	1	0.91210	0.20319	20.1498	<.0001	2.490	1.672	3.707
AGE4	1	1.22998	0.19741	38.8184	<.0001	3.421	2.323	5.037
AGE5	1	1.82465	0.19486	87.6867	<.0001	6.201	4.232	9.084
BP2	1	0.23912	0.14112	2.8712	0.0902	1.270	0.963	1.675
BP3	1	0.60403	0.14556	17.2192	<.0001	1.829	1.375	2.433
EXSMOK	1	0.20508	0.16452	1.5538	0.2126	1.228	0.889	1.695
CUSMOK	1	0.52770	0.13815	14.5914	0.0001	1.695	1.293	2.222
BMI2	1	-0.43098	0.34115	1.5959	0.2065	0.650	0.333	1.268
BMI3	1	-0.65933	0.34171	3.7230	0.0537	0.517	0.265	1.010

2) Poisson 회귀모형

: Poisson 회귀분석을 위해 4가지 변수의 조합수는 135개 임.

사용한 변수 SDCD는 각 조합에서 발생한 총 사망자 수이고, LSSDAYS는 그 조합에서 관찰한 총 인일(person-days)을 자연대수한 값이다.

```
PROC GENMOD DATA=POP3 DESCENDING;
     MODEL SDCD=AGE2 AGE3 AGE4 AGE5 BP2 BP3 EXSMOK CUSMOK BMI2 BMI3
                /DIST=POISSON
                 LINK=LOG
                 OFFSET=LSSDAYS;
RUN;
```

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	-12.2584	0.4016	-13.0454	-11.4714	931.91	<.0001
AGE2	1	0.5661	0.2112	0.1521	0.9801	7.18	0.0074
AGE3	1	0.9106	0.2032	0.5123	1.3089	20.08	<.0001
AGE4	1	1.2275	0.1974	0.8405	1.6144	38.65	<.0001
AGE5	1	1.8187	0.1949	1.4367	2.2006	87.09	<.0001
BP2	1	0.2380	0.1411	-0.0386	0.5146	2.84	0.0917
BP3	1	0.6011	0.1456	0.3159	0.8864	17.06	<.0001
EXSMOK	1	0.2044	0.1645	-0.1180	0.5269	1.54	0.2140
CUSMOK	1	0.5255	0.1381	0.2548	0.7963	14.47	0.0001
BMI2	1	-0.4293	0.3412	-1.0981	0.2394	1.58	0.2083
BMI3	1	-0.6553	0.3418	-1.3252	0.0145	3.68	0.0552
Scale	0	1.0000	0.0000	1.0000	1.0000		

3) Nested case-control 자료에서의 회귀분석

: 본 자료분석을 위해 각 case에 대해 3배의 control (1:3 matching)을 추출하였다.

```
DATA CASECON; SET KDR.A10;
   /* PEARCE N. Incidence density matching with a simple SAS computer program.
      위의 논문을 수정하여 survival time에 기초한 Sampling을 하였음 */
RUN;

PROC PHREG DATA=CASECON;
     STRATA CASESET;
     MODEL NCASE=AGE2 AGE3 AGE4 AGE5 BP2 BP3 EXSMOK CUSMOK BMI2 BMI3 / TIES=DISCRETE RL;
RUN;
```

- Incidence density sampling을 한 자료의 구조는 다음과 같다.

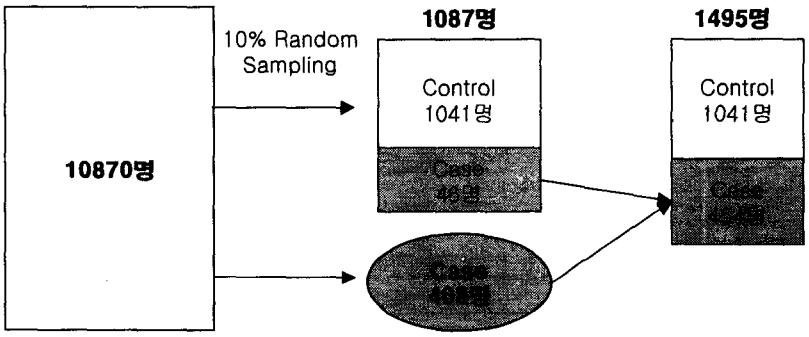
Obs	DCD	SDAYS	CASESET	gAGE	gBP	gSMOK	gBMI	Ncase
1	1	8	1	3	1	3	2	1
2	0	2921	1	1	1	3	2	2
3	0	2921	1	4	2	1	3	2
4	0	2921	1	2	3	3	3	2
5	1	21	2	4	3	3	3	1
6	0	2921	2	3	2	3	2	2
7	0	2921	2	4	3	1	3	2
8	0	2921	2	4	2	1	3	2
9	1	24	3	5	3	3	3	1
10	0	2921	3	1	2	1	3	2
11	0	2921	3	2	2	3	2	2
12	1	1373	3	4	2	3	2	2
13	1	23	4	2	2	1	1	1
14	0	2921	4	5	2	.	1	2
15	0	2921	4	2	2	3	3	2
16	0	2921	4	3	3	1	3	2

이하 계속

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits	
AGE2	1	0.56902	0.23318	5.9548	0.0147	1.767	1.119	2.790
AGE3	1	0.80083	0.23032	12.0895	0.0005	2.227	1.418	3.498
AGE4	1	1.28973	0.22616	32.5220	<.0001	3.632	2.331	5.658
AGE5	1	1.87689	0.22914	67.0932	<.0001	6.533	4.169	10.237
BP2	1	0.15854	0.16699	0.9014	0.3424	1.172	0.845	1.626
BP3	1	0.47897	0.17792	7.2471	0.0071	1.614	1.139	2.288
EXSMOK	1	0.15995	0.19841	0.6499	0.4202	1.173	0.795	1.731
CUSMOK	1	0.46445	0.16689	7.7447	0.0054	1.591	1.147	2.207
BMI2	1	-0.91115	0.52537	3.0078	0.0829	0.402	0.144	1.126
BMI3	1	-1.15164	0.52689	4.7775	0.0288	0.316	0.113	0.888

4) Case-cohort 자료분석

: 본 자료분석에서 부분 코호트는 전체코호트에서 10%를 랜덤하게 추출하였다. 그 과정은 다음과 같다.




```
DATA TEMP; SET KDR.A10;
    /* Therneau T, Li H. Computing the Cox Model for Case Cohort Design. 의 논문에서
     제시한 방법(Self & Prentice 방법)을 이용하여 자료를 구축*/
RUN;
PROC PHREG DATA=TEMP;
  MODEL SDAYS*DCD(0)=AGE2 AGE3 AGE4 AGE5 BP2 BP3 EXSMOK CUSMOK BMI2 BMI3
                    / OFFSET=DUMMY RL;
RUN;
```

Obs	DCD	SDAYS	gAGE	gBP	gSMOK	gBMI	SC10	CASECON	DUMMY
1	1	2682	5	3	3	2	.	.	-100
2	0	2921	5	3	1	3	1	0	0
3	1	1443	5	1	2	3	1	1	-100
4	0	1443	5	1	2	3	1	1	0
5	0	2921	5	3	1	2	1	0	0
6	1	1842	5	3	2	2	.	.	-100
7	1	2702	5	3	1	3	.	.	-100
8	1	58	5	2	3	2	.	.	-100
9	0	2921	5	2	1	2	1	0	0

이하 계속

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits	
AGE2	1	0.62341	0.21170	8.6722	0.0032	1.865	1.232	2.824
AGE3	1	0.96284	0.20409	22.2563	<.0001	2.619	1.756	3.907
AGE4	1	1.29820	0.19742	43.2427	<.0001	3.663	2.487	5.393
AGE5	1	1.80281	0.19556	84.9821	<.0001	6.067	4.135	8.900
BP2	1	0.19974	0.14160	1.9899	0.1584	1.221	0.925	1.612
BP3	1	0.63439	0.14673	18.6942	<.0001	1.886	1.415	2.514
EXSMOK	1	0.22267	0.16494	1.8224	0.1770	1.249	0.904	1.726
CUSMOK	1	0.57234	0.13904	16.9445	<.0001	1.772	1.350	2.328
BMI2	1	-0.19552	0.34341	0.3242	0.5691	0.822	0.420	1.612
BMI3	1	-0.43485	0.34407	1.5972	0.2063	0.647	0.330	1.271

이상 분석의 위험비(RR 또는 OR)를 정리하면 다음과 같다.

독립변수	범주	전체 자료		Nested Case-Control		Case-Cohort	
		Cox 모형	포아송모형	A 방법[1]	B 방법[2]	C 방법[3]	D 방법[4]
연령	40세 미만	1	1	1	1	1	1
	40-44	1.76	1.76	1.77	1.73	1.87	1.79
	45-49	2.49	2.49	2.23	2.25	2.62	2.34
	50-54	3.42	3.41	3.63	3.55	3.66	3.17
	55세 이상	6.20	6.16	6.53	6.54	6.07	4.47
혈압	Normal	1	1	1	1	1	1
	Prehypertension	1.27	1.27	1.17	1.16	1.22	1.18
	Hypertension	1.83	1.82	1.61	1.64	1.89	1.66
흡연	비흡연	1	1	1	1	1	1
	과거흡연	1.23	1.23	1.17	1.03	1.25	1.22
	현재흡연	1.70	1.69	1.59	1.51	1.77	1.61
비만도	18.5 미만	1	1	1	1	1	1
	18.5-23.0	0.65	0.65	0.40	0.32	0.82	0.78
	23 이상	0.52	0.52	0.32	0.25	0.65	0.68

1) A 방법은 matching 된 자료의 정보를 이용한 discrete-time에 대한 stratified Cox 모형

2) B 방법은 matching 정보를 이용하지 않은 logistic regression

3) Therneau & Li 의 논문에서 설명한 Self & Prentice 방법으로 회귀계수를 추정

4) Self & Prentice 방법을 이용하지 않고 단순히 Case-cohort에서 Cox 회귀모형을 수행

이상에서 추적관찰된 자료에서의 여러 가지 회귀분석 방법들을 살펴보았다. 한편 회귀모형의 구축할 때 모형의 기본가정, 모형의 적합성 등을 검토하는 것도 매우 중요하나 본 강의에서는 약하기로 한다.