관상동맥질환(CAD)의
연관성 연구 및 예측 모형 구축

과목 : 회귀분석 2

교수 : 조상훈 교수님

학과 : 정보통계보험수리학과

조장 : 20171421 김성연

조원 : 20171424 김예지,

20171455 조경덕

제출일자 : 2018 년 12 월 18 일

목차

Ⅰ. 서론

1. 연구의 목적 및 필요성

요즘 사회는 최첨단 기술과 발달로 인해 대체적으로 사람들의 삶의 질이 향상되었습니다. 하지만 질병은 여전히 발생하고 있고, 최첨단 기술도 어찌하지 못하는 질병도 존재합니다. 우리가 최근에 와서야 통계가 질병에 도움을 준다고 생각하지만, 사실 예전에도 통계로 사람들의 질병을 어느 정도 예상할 수 있었습니다. 그 예로 나이팅게일을 들 수 있습니다. 나이팅게일은 전쟁터의 병사들의 사망원인을 담은 도표와 자료 분석등을 통해 사망률을 40%에서 2%로 줄이는 놀라운 일을 행했습니다. 그 당시에서 오랜 시간이 지난 현재는 더 많은 설명변수들과 통계 프로그램을 통해 질병의 원인을 더 체계적으로 연구할 수 있게 되었습니다. 이번에 우리가 알아볼 내용은 나이 성 흡연여부 고혈압 헤모글로빈 등의 총 12 가지의 설명변수를 통해 관상동맥질환과의 연관성을 연구해보려고 합니다.

40'연구"방법

본 분석에서는 반응 변수 CADGROUP (관상동맥질환, 0: 심각하지 않음; 1: 심각한 상태), 와 설명변수 age (나이), Sex (성, 1: 여성; 2: 남성), smoking (흡연 여부, 0: 비흡연; 1: 흡연), HTN (고혈압, 0: 정상혈압; 1:고혈압), TCHOL (총 콜레스테롤), hdl (HDL 콜레스테롤), LDL (LDL 콜레스테롤), hemoglobin (헤모글로빈), WBC (백혈구), neutrophil (호중구), lymphocyte (림프구), DM (제 2 형 당뇨병, 0: 정상; 1: 당뇨환자)의 데이터를 사용했습니다. 반응변수 CADGROUP(관상동맥질환)이 0 과 1 로 이루어진 이진모형이기 때문에 로지스틱 회귀분석을 사용하도록 합니다.

Ⅱ. 본론

1. 설명변수에 대한 탐구

1-1. '나이'와 '백혈구' 변수에 대해

'나이' 변수에 대해 관찰해보면 모든 질병에서 나이가 높을수록 발병 확률 또한 증가한다는 사실은 대부분 알고 있을 것입니다. 나이가 들면 근육이 줄어들고, 근육이 줄면 각종 질병의 위험이 올라가기 때문입니다. 그러므로 t 통계량 값과 상관없이 '나이'변수는 설명변수로서 넣기로 결정합니다.

또한 관상동맥우회술을 시행한 관상동맥질환 환자에서 수술 전 백혈구 수치와 수술 후 경과의 관계[1]를 주제로 한 논문에 따르면 '여러 연구에서 백혈구 수치의 증가가 심혈관 질환에 의한 사망에 영향을 준다는 결과를 보고하고 있다'고 합니다. 이를 통해 '백혈구' 변수 또한 설명변수에 넣기로 결정합니다.

1-2. '콜레스테롤' 변수에 대해

다음으로는 '콜레스테롤' 변수에 대해서 분석해보기로 했습니다. 우선 총 콜레스테롤[2] 은 네이버 지식 백과에 따르면 HDL 콜레스테롤과 LDL 콜레스테롤, 그리고 (중성지방)/5 의 합으로 이루어져 있습니다.

일반적으로 총 콜레스테롤이 높을수록 관상동맥질환 발병률이 증가한다고 알려져 있는데, 이는 '총 콜레스테롤' 변수가 관상동맥질환에 영향을 끼치는 것이 아니라 'LDL 콜레스테롤' 변수가 관상동맥질환에 영향을 끼치는 것을 대신 설명했다고 볼 수 있습니다.

또한 학기술인공제회에서 발췌한 내용에 따르면 HDL 콜레스테롤은 혈관에 쌓인 불필요한 LDL 콜레스테롤을 간으로 돌려보내기 때문에, 관상동맥질환 발병률을 낮춰주고, LDL 콜레스테롤이 과다하게 있을 때 혈관에 흡착하여 쌓이기 때문에 질환 발병률을 높입니다.

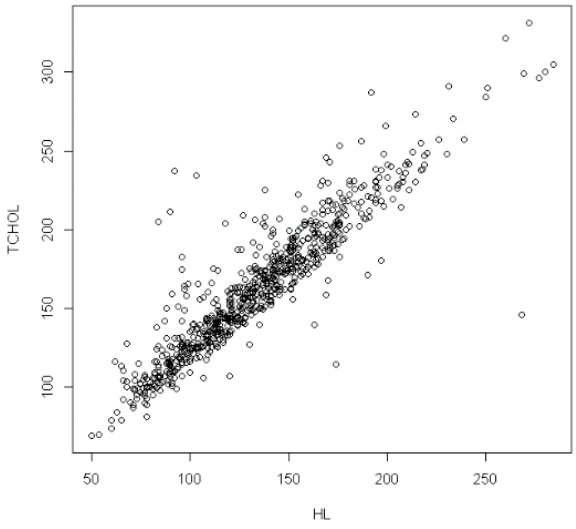
하지만, HDL 이 무조건 좋은 수치고, 중요한 수치라고 말하기 힘듭니다. 허핑턴포스트 뉴스 기사 '좋은 콜레스테롤(HDL)이 무조건 좋은 건 아니다(연구결과)'[3] 에 따르면 "의료계에서는 높은 HDL 콜레스테롤 수치가 근본적으로 심장 보호 역할을 하는데 도움이 된다고 오랫동안 믿어왔다. 그리고 그런 인식을 반영하듯 제약회사들은 좋은 콜레스테롤 수치를 높이는 약 개발에 열중했는데 정작 효과는 미미했다. 이번 연구는 오히려 유전적 돌연변이를 지닌 선천적으로 좋은 콜레스테롤 수치가 높은 사람들의 심장질환 확률이 더 높다고 밝혔다."라는 내용이 있는데, 이에 따르면 HDL 콜레스테롤에 초점을 맞추기보다는 LDL 콜레스테롤에 초점을 맞추어 설명을 하는 것이 더 적합하다 생각했습니다.

---

1 https://dspace.inha.ac.kr/bitstream/10505/19169/1/20105.pdf

2 https://terms.naver.com/entry.nhn?docId=3535926&cid=58572&categoryId=58572

3 https://www.huffingtonpost.kr/2016/03/18/story_n_9493302.html




```
> cor(TCHOL,HL)
[1]  0.9159464
```

지금까지의 가정을 확인하기 위해, 주어진 표본으로 분석을 해보면 HDL 콜레스테롤 + LDL 콜레스테롤 값과 총 콜레스테롤 (TCHOL) 사이에 선형관계가 나타나는 것을 볼 수 있습니다. 또한 총 콜레스테롤과 HL(HDL 콜레스테롤 + LDL 콜레스테롤) 간 상관계수가 약 0.916 로 매우 높게 나타납니다. 즉, 총 콜레스테롤이 HDL 콜레스테롤 + LDL 콜레스테롤에 선형관계가 있음을 알 수 있습니다.

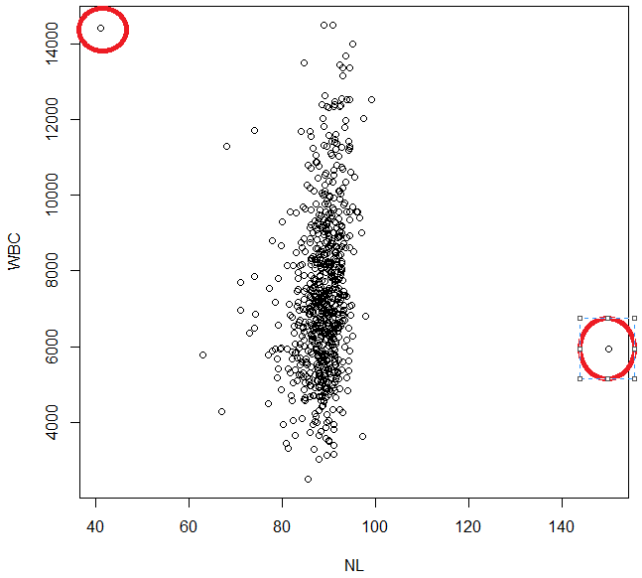
이를 바탕으로 우선 주어진 표본을 바탕으로 총 콜레스테롤이 HDL 콜레스테롤과 LDL 콜레스테롤의 합보다 지나치게 큰 경우(총 콜레스테롤 – HDL 콜레스테롤 – LDL 콜레스테롤 > 80, 중성지방이 400 보다 크다고 나온 결과) 잘못 측정된 표본이라고 생각하고 표본에서 제외하겠습니다. 그 결과 13, 20, 36, 52, 64, 129, 176, 229, 348, 357, 447, 450, 521, 542, 642, 681, 736 번째 데이터를 표본에서 제외하게 되었습니다.

그리고 다중공선성을 피하기 위해, 변수를 하나 제거해야 하는데, 앞서 살펴본 이론에 따르면 '총 콜레스테롤' 변수보단 'LDL 콜레스테롤' 변수가 실제로 관상동맥질환에 연관이 있음을 보였기 때문에 '총 콜레스테롤' 변수를 제거하기로 결정합니다.

1-3 '호중구', '림프구' 변수에 대해

학회지 '급성 뇌경색 환자에서 입원시 호중구/림프구 비가 예후에 미치는 영향[4]에 따르면 "특히, 심혈관 질환 환자에서 백혈구 수치와 그 아형이 사망 또는 심근경색의 예측인자로서 유용하여 특히 호중구/림프구 비의 유의성이 가장 높았고, 이는 기존연구에서 보고한 CRP 의 예측도보다 높았다."라는 내용을 찾아볼 수 있습니다.

[4] http://jkna.org/upload/pdf/201003005.pdf




```
> cor(neutrophil,lymphocye)
 [1] -0.8875041
```

또한 (호중구 + 림프구) 가 위 그림을 보아 일정한 값에 수렴함을 알 수 있습니다. 또한 호중구와 림프구의 상관계수가 약 -0.888 로 강한 상관관계를 보입니다. 우선 빨간색 동그라미 표시에 표본을 특이 값으로 생각하여 표본에서 제외하겠습니다. 그 결과 201, 502 번째 데이터를 표본에서 제외하게 되었습니다.

또 호중구+림프구 값이 일정한 값에 수렴한다는 것은 다중공선성 문제를 야기합니다. 그래서 다중공선성을 피하기 위해 변수 하나를 제거해야 합니다. 앞에 자료에 따르면 호중구/림프구 비를 사용하는 것이 상당히 효과가 있다는 점을 제시하였기에 변수 하나를 제거하기 보다는 호중구/림프구 값을 설명변수로 사용하려 합니다.

```
> N_L
[64]  1.2411168  2.7991453  4.4252874  1.2626263  1.6767372
[71]  2.8750000  5.1283784  [illegible]  2.3071161
[78]  3.4619289  1.8437500 15.3684211 29.9000000  1.8229814
[85]  1.6617647  3.0407240  [illegible]  3.5072464
[92]  1.4385475  1.6852941  3.0518868  7.5727273  1.5813253
[99]  1.6407186  1.8126984  1.6695652  2.7008929  1.2356021
106]  4.9910600  3.1838565  0.9766454  2.1290323  2.7000000
113]  1.3273196  1.3437500  6.2031250  1.7243590  2.4749035
```

다만 위 그림과 같이 호중구/림프구 비는 분모가 작을 때 값이 급격하게 커질 수 있는데, 다른 값에 비해 월등히 큰 값을 가진 변수가 있다면 영향력 관측치가 될 수 있습니다. 이런 현상을 방지하기 위해 호중구/림프구 비를 사용할 때 자연로그변환을 사용하기로 결정합니다.

2. 회귀분석 로지스틱 실시

지금까지 변수를 살펴본 것을 토대로 로지스틱 회귀분석을 실시하겠습니다. 우선 앞에서 설명변수에 대해 분석하였습니다. 그 결과 '나이' 변수, '백혈구' 변수, 'HDL 콜레스테롤' 변수, 'LDL

콜레스테롤’ 변수, ‘log(호중구/림프구)’ 변수 총 다섯 개를 설명변수로 사용하기로 하였고, ‘성별’ 변수, ‘흡연여부’ 변수, ‘제 2 형 당뇨병 유무’ 변수, ‘헤모글로빈’ 변수, ‘고혈압 유무’ 변수 총 다섯 개 설명변수 후보군으로 사용하기로 결정했습니다.

사용해야 하는 다섯 개의 설명변수가 존재하고, 나머지 다섯 개의 설명변수는 사용할 지에 대해 유무를 확인해야 하기 때문에, 사용해야 하는 다섯 개의 설명변수를 초기모형으로 하여 단계별 회귀를 실시하였습니다.

```
fit <- glm(formula = CADGROUP ~ age + hdl + LDL + log(N_L)+ WBC,
 family = binomial(link = "logit"), data = data, trace = TRUE)
fit2 <- step(fit, scope = ~ age + hdl + LDL + log(N_L)+ WBC+
Sex+smoking+HTN+hemoglobin+DM,direction="both")
```

```
             Df Deviance    AIC
<none>            729.23 747.23
+ LDL         1   727.54 747.54
- hdl         1   732.32 748.32
```

```
Call:
glm(formula = CADGROUP ~ age + hdl + log(N_L) + WBC + smoking +
    Sex + DM + hemoglobin, family = binomial(link = "logit"),
    data = data, trace = TRUE)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9438  -0.8037  -0.4950   0.8499   2.5817

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.678e+00  1.177e+00  -1.425  0.15402
age          3.298e-02  8.479e-03   3.890  0.00010 ***
hdl         -1.484e-02  8.582e-03  -1.729  0.08384 .
log(N_L)     5.020e-01  1.848e-01   2.716  0.00660 **
WBC          8.776e-05  4.758e-05   1.845  0.06508 .
smokingTRUE  2.015e+00  2.698e-01   7.467 8.20e-14 ***
Sexman      -1.714e+00  2.727e-01  -6.286 3.26e-10 ***
DMTRUE       6.014e-01  2.069e-01   2.906  0.00366 **
hemoglobin  -1.156e-01  6.312e-02  -1.832  0.06695 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 897.23  on 725  degrees of freedom
Residual deviance: 729.23  on 717  degrees of freedom
AIC: 747.23

Number of Fisher Scoring iterations: 5
```

[그림 1]

적합결과, 예상과 다르게 반드시 넣어야 하는 설명변수인 'LDL 콜레스테롤' 변수가 마지막 부분에서 설명변수로 선택 되지 못했습니다. 'LDL 콜레스테롤' 변수는 앞서 말한 바에 의하면 반드시 들어가야 하는 설명변수입니다. 이 변수가 선택 되지 못했다는 것은 두 가지로 설명할 수 있습니다. 첫 번째는 이 설명변수가 정말로 반응변수를 설명하지 못해서이고, 두 번째는 'LDL 콜레스테롤' 변수가 설명할 부분을 다른 변수가 대신 설명하는 경우입니다.

앞에 설명한 'LDL 콜레스테롤에 초점을 맞춘다'는 기준을 토대로, 두 번째 경우라고 생각하고 다시 분석하겠습니다. 두 번째 경우가 옳다면, 다른 설명변수가 'LDL 콜레스테롤' 변수를 설명해야 합니다. 그래야 위에 로지스틱 회귀모형에서 다른 설명변수가 'LDL 콜레스테롤' 변수를 대신해서 반응변수를 설명하기 때문입니다. 그래서 'LDL 콜레스테롤' 변수를 반응변수로, 위 로지스틱 회귀모형에 다른 변수를 설명변수로 놓고 회귀분석을 실시했습니다.

```
Call:
lm(formula = LDL ~ age + hdl + log(N_L) + WBC + smoking + Sex +
    DM + hemoglobin, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-72.967 -23.162  -2.084  18.723 151.213

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.410e+01  1.559e+01  -0.905  0.36587
age          1.965e-02  1.026e-01   0.192  0.84811
hdl          3.342e-01  1.032e-01   3.239  0.00125 **
log(N_L)     4.731e-01  2.477e+00   0.191  0.84858
WBC          5.554e-04  6.396e-04   0.868  0.38553
smokingTRUE -6.291e+00  3.177e+00  -1.980  0.04807 *
Sexman       1.090e+01  3.082e+00   3.538  0.00043 ***
DMTRUE      -6.754e+00  2.963e+00  -2.279  0.02294 *
hemoglobin   6.123e+00  8.654e-01   7.075 3.56e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 32.32 on 717 degrees of freedom
Multiple R-squared:  0.1203,    Adjusted R-squared:  0.1105
F-statistic: 12.26 on 8 and 717 DF,  p-value: < 2.2e-16
```

예상대로, 'HDL 콜레스테롤' 변수, '성별' 변수, '흡연여부' 변수, '제 2 형 당뇨병 유무' 변수, '헤모글로빈' 변수 총 5 개변수가 'LDL 콜레스테롤' 변수를 유의하게 설명하는 것을 볼 수 있습니다.

[그림 1]에서 유의수준 0.05 를 기준으로 p 값이 크며 연속형 설명변수인 'HDL 콜레스테롤' 변수, '헤모글로빈' 변수는 모형에서 제거합니다. 그리고 p 값이 상당히 작은 '성별' 변수, '흡연여부' 변수, '제 2 형 당뇨병 유무' 변수는 쉽게 제거하기 힘듭니다. 또한 이산형 설명변수기 때문에 'LDL 콜레스테롤' 변수에 대한 설명력이 크지 않다고 판단합니다. 그래서 이 세 변수는 제거하기보단, 'LDL 콜레스테롤' 변수와 교호작용항 변수을 추가한 채로 모형을 새우려고 합니다.

다시 말하면 '나이' 변수, '백혈구' 변수, 'LDL 콜레스테롤' 변수, 'log(호중구/림프구)' 변수, '성별' 변수, '흡연여부' 변수, '제 2 형 당뇨병 유무' 변수, 'LDL 콜레스테롤' : '성별' 변수, 'LDL 콜레스테롤' : '흡연 여부' 변수, 'LDL 콜레스테롤' : '제 2 형 당뇨병 유무' 변수를 설명변수로 하여 회귀분석을 실시합니다.

```
Call:
glm(formula = CADGROUP ~ age + log(N_L) + WBC + smoking + Sex +
    DM + LDL + LDL:Sex + LDL:DM + LDL:smoking, family = binomial(link = "logit"),
    data = data, trace = TRUE)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9830  -0.7962  -0.4835   0.8577   2.8589

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)     -3.812e+00  7.527e-01  -5.065 4.09e-07 ***
age              3.999e-02  8.282e-03   4.828 1.38e-06 ***
log(N_L)         5.349e-01  1.829e-01   2.924  0.00345 **
WBC              7.290e-05  4.727e-05   1.542  0.12301
smokingTRUE      1.458e+00  7.372e-01   1.978  0.04790 *
Sexman          -1.011e-01  6.997e-01  -0.145  0.88507
DMTRUE          -4.067e-01  5.809e-01  -0.700  0.48383
LDL             -4.622e-03  4.049e-03  -1.141  0.25372
Sexman:LDL      -1.859e-02  8.083e-03  -2.300  0.02145 *
DMTRUE:LDL       1.327e-02  6.765e-03   1.961  0.04988 *
smokingTRUE:LDL  6.116e-03  8.448e-03   0.724  0.46914
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 897.23  on 725  degrees of freedom
Residual deviance: 722.93  on 715  degrees of freedom
AIC: 744.93

Number of Fisher Scoring iterations: 5
```

'성별' 변수, ''제 2 형 당뇨병 유무' 변수와 'LDL 콜레스테롤' 변수간 교호작용항이 유의합니다. 이는 성별과 제 2 형 당뇨병 유무가 바뀌면 'LDL 콜레스테롤' 변수의 회귀계수가 바뀐다는 것을 의미합니다. 또한 '성별' 변수, ''제 2 형 당뇨병 유무' 변수의 P 값이 상당히 증가했습니다. 이는 성별과 제 2 형 당뇨병 유무가 실제로 관상동맥질환에 영향을 끼쳤다기보다는 'LDL 콜레스테롤' 회귀계수를 조정해주는 역할로 볼 수 있습니다.

또 '흡연 여부' 변수와 'LDL 콜레스테롤' 변수간 교호작용항은 유의하지 않습니다. 그래서 '흡연 여부' 변수와 'LDL 콜레스테롤' 변수간 교호작용항 변수, '성별' 변수, '제 2 형 당뇨병 유무' 변수는 제거하고 다시 회귀분석을 하겠습니다. 다만 'LDL 콜레스테롤' 변수는 제거하면 회귀계수를 해석하는데 어렵기 때문에 유지한 채로 회귀분석을 하기로 결정했습니다.

그 결과 최종으로 적합 된 로지스틱 회귀모형이 나오게 됩니다. [그림 1] 회귀모형에 비해 AIC 가 약 7 정도 감소된 좋은 모형임을 알 수 있습니다.

```
Call:
glm(formula = CADGROUP ~ age + log(N_L) + WBC + smoking + LDL +
    LDL:Sex + LDL:DM, family = binomial(link = "logit"), data = data,
    trace = TRUE)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0533  -0.7993  -0.4909   0.8489   2.7906

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.118e+00  6.961e-01  -5.916 3.30e-09 ***
age          4.047e-02  8.231e-03   4.917 8.79e-07 ***
log(N_L)     5.228e-01  1.828e-01   2.860 0.004231 **
WBC          7.473e-05  4.721e-05   1.583 0.113445
smokingTRUE  1.907e+00  2.576e-01   7.404 1.32e-13 ***
LDL         -1.794e-03  3.027e-03  -0.593 0.553454
LDL:Sexman  -1.876e-02  2.829e-03  -6.631 3.33e-11 ***
LDL:DMTRUE   8.874e-03  2.374e-03   3.738 0.000186 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 897.23  on 725  degrees of freedom
Residual deviance: 724.28  on 718  degrees of freedom
AIC: 740.28

Number of Fisher Scoring iterations: 5
```

3. 결정된 회귀모형 진단

3-1 다중공선성 진단

```
vif(comfit2)
      age    log(N_L)         WBC smokingTRUE         LDL  LDL:Sexman  LDL:DMTRUE
 1.160845    1.180780    1.168040    1.602003    1.108907    1.736078    1.035690
```

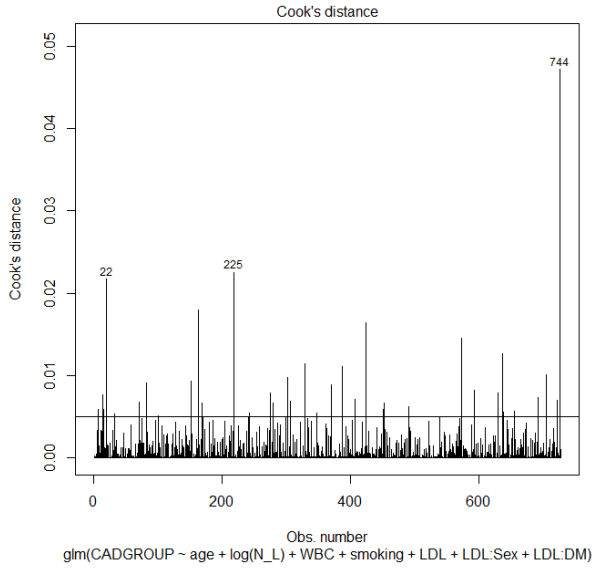
모든 설명변수들의 vif 값이 10 이하이므로 다중공선성이 존재하지 않는다고 할 수 있다.

3-2 영향력 관측치 진단

아래 왼쪽 그림은 쿡의 거리 통계량으로 영향력 관측치를 측정하는 값 입니다. 쿡의 거리 통계량에서 영향력 관측치임을 나타내는 기준은 3.67/(728-7) = 약 0.005 로 그래프 위에 선을 표시하였습니다.

아래 오른쪽 그림은 'influence.measures()' 함수를 사용하여 영향력 관측치 진단한 것 입니다. 이를 기준으로 보면 7, 15, 22, 50, 86, 111, 134, 149, 170, 175, 211, 212, 225, 271, 292, 311, 338, 358, 380, 413, 418, 560, 587, 593, 640, 663, 744 번째 표본이 영향력 관측치 임을 알 수 있습니다.

영향력 관측지로 분류된 표본을 제외한 모형을 최종 모형으로 세우기로 합니다. 이 표본들은 대부분 앞서 제거하지 못한 '백혈구' 변수의 높은 지렛점으로 추측됩니다. 실제로 다음 페이지에 나오는 최종 모형을 살펴보면 '백혈구' 변수의 p 값이 눈에 띄게 줄어 든 것이 확인됩니다.



	dfb.LDL.D	dffit	cov.r	cook.d	hat	inf
1	0.018832	-0.06043	1.012	2.63e-04	0.006865	
2	0.006606	-0.03220	1.013	6.94e-05	0.004401	
3	0.011488	-0.04279	1.012	1.27e-04	0.004884	
4	0.011831	-0.08002	1.015	4.70e-04	0.010658	
5	-0.032554	0.13010	0.972	3.31e-03	0.004293	
6	0.107997	0.13217	1.004	1.68e-03	0.010861	
7	-0.173938	-0.24704	1.030	5.85e-03	0.036173	*
8	-0.056798	-0.12836	1.010	1.45e-03	0.012650	
9	0.001467	-0.00841	1.013	4.50e-06	0.002261	
10	-0.018397	0.08592	1.014	5.56e-04	0.010486	
11	0.000889	-0.00770	1.013	3.77e-06	0.002215	
12	0.081760	0.15609	0.989	3.30e-03	0.008729	
14	0.135363	0.18066	1.010	3.26e-03	0.018407	
15	0.017823	0.26858	1.025	7.63e-03	0.035081	*
16	0.018357	-0.05304	1.009	2.07e-04	0.004607	
17	-0.043255	0.23685	1.018	5.88e-03	0.028127	
18	0.080383	0.12588	1.019	1.25e-03	0.017614	
19	0.012636	-0.06558	1.015	3.05e-04	0.009089	
21	-0.037218	0.09358	0.985	1.15e-03	0.003288	
22	0.168777	0.32469	0.990	2.17e-02	0.024662	*
23	-0.020561	0.13316	1.011	1.56e-03	0.013759	
24	-0.042504	-0.10216	1.007	8.94e-04	0.008813	
25	-0.028687	0.13144	1.010	1.52e-03	0.013297	
26	0.003271	-0.02991	1.017	5.83e-05	0.007030	
27	-0.023160	0.13783	1.007	1.78e-03	0.012390	
28	0.011757	-0.05581	1.021	2.09e-04	0.012027	
29	0.007276	-0.02562	1.015	4.27e-05	0.005192	
30	-0.031620	-0.06498	1.017	2.94e-04	0.010522	

4. 최종 모형 적합 및 해석

4-1. 최종 모형 적합.

```
Call:
glm(formula = CADGROUP ~ age + log(N_L) + WBC + smoking + LDL +
    LDL:Sex + LDL:DM, family = binomial(link = "logit"), data = ddata,
    trace = TRUE)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2192  -0.7616  -0.4250   0.7435   2.2677

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.9733034  0.7844633  -6.340 2.30e-10 ***
age          0.0508479  0.0092595   5.491 3.99e-08 ***
log(N_L)     0.3387356  0.2014836   1.681  0.09272 .
WBC          0.0001263  0.0000508   2.486  0.01291 *
smokingTRUE  2.2086596  0.2914949   7.577 3.54e-14 ***
LDL         -0.0021582  0.0034447  -0.627  0.53097
LDL:Sexman  -0.0238965  0.0033830  -7.064 1.62e-12 ***
LDL:DMTRUE   0.0083555  0.0027457   3.043  0.00234 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 854.22  on 700  degrees of freedom
Residual deviance: 661.30  on 693  degrees of freedom
AIC: 677.3
```

$$\ln\left(\frac{\pi}{1-\pi}\right) = -4.9733 + 0.0608\,age - 0.0026\,LDL - 0.0239\,LDL \times Sex + 0.0084\,LDL \times DM + 0.0001\,WBC + 0.3387\log\left(\frac{neutrophil}{lymphocyte}\right) + 2.2087\,smoking$$

(단, 남성일 때 Sex 가 1, 여성일 때 Sex 가 0, 흡연자일 때 smoking 이 1, 비흡연자일 때 smoking 이 0, 당뇨환자일 때 DM 이 1, 정상일 때 DM 이 0, 파이는 관상동맥질환 발병률 의미)

나이가 1 세 증가할 때 관상동맥질환이 발생할 오즈가 1.0521 변화합니다.

제 2 종 당뇨병 환자인 남자가 LDL 이 1 단위 증가할 때 관상동맥질환이 발생할 오즈가 0.9821 변화합니다.

제 2 종 당뇨병 환자인 여자가 LDL 이 1 단위 증가할 때 관상동맥질환이 발생할 오즈가 1.0058 변화합니다.

제 2 종 당뇨병 환자가 아닌 남자가 LDL 이 1 단위 증가할 때 관상동맥질환이 발생할 오즈가 0.9738 변화합니다.

제 2 종 당뇨병 환자가 아닌 여자가 LDL 이 1 단위 증가할 때 관상동맥질환이 발생할 오즈가 0.9974 변화합니다.

백혈구가 1 단위 증가할 때 관상동맥질환이 발생할 오즈가 1.0001 변화합니다.

로그(호중구/림프구) 값이 1 단위 증가할 때 관상동맥질환이 발생할 오즈가 1.4031 변화합니다.

비흡연자에 비해 흡연자가 관상동맥질환이 발생할 오즈가 9.1039 변화합니다.

4-2. 최종 모형 해석

오즈는 1 보다 크다면 관상동맥질환 발병률을 증가시킨다고 해석되고, 1 보다 작다면 관상동맥 질환 발병률을 감소시킨다고 해석할 수 있습니다. 이를 토대로 최종적으로 적합 된 회귀식을 해석하겠습니다.

앞서 설명변수 분석에서 나이가 들면 어느 질환이든 발병률이 높아져 '나이' 변수를 넣었었습니다. 예상대로 나이가 1 세 증가할 때 오즈의 변화량이 1 보다 크기 때문에, 관상동맥질환에서도 나이가 들면 발병률이 유의하게 높아짐을 알 수 있습니다.

호중구/림프구 비도 최종모형에서는 p 값이 조금 크지만, 영향력 관측치를 제외하기 전 회귀모형에서 로그(호중구/림프구) 변수가 충분히 유의함을 보였습니다. 또한 로그(호중구/림프구)가 1 단위 증가할 때 오즈의 변화량이 1 보다 크기 때문에 '호중구/림프구' 변수가 증가하면 발병률이 유의하게 높아짐을 알 수 있습니다.

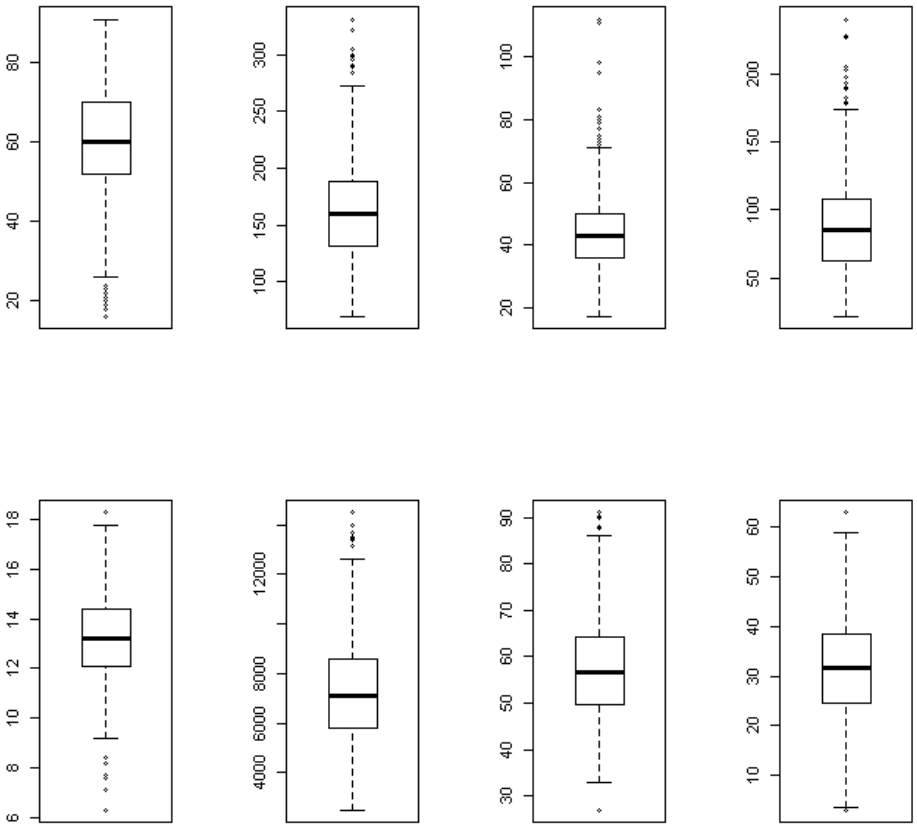
앞서 설명변수 분석에서 백혈구 수치가 커지면 관상동맥질환에 발병률이 높아진 다는 자료를 보고 '백혈구' 변수를 보았습니다. 예상대로 백혈구가 1 단위 증가할 때, 오즈의 변화량이 1 보다 크기 때문에 '백혈구' 변수가 증가하면 발병률이 유의하게 높아집니다. 그리고 '백혈구' 변수의 p 값은 충분히 유의하지만 회귀계수 적합값이 약 0.0001 로 다소 낮은데, 이는 '백혈구' 변수의 값 자체가 다른 설명변수의 값보다 월등히 크기 때문에 벌어지는 일 입니다.

비흡연자의 비해 흡연자에 오즈의 변화량이 1 보다 월등히 큽니다. 이는 비흡연자에 비해 흡연자의 관상동맥질환 발병률이 유의미하게 높아짐을 알 수 있습니다.

제 2 종 당뇨병 여성환자가 LDL 콜레스테롤이 1 단위 증가할 때 오즈가 1 보다 크기 때문에 이 경우에 'LDL 콜레스테롤' 변수의 증가는 발병률을 유의하게 높인다고 볼 수 있습니다. 하지만 그 이외의 경우에는 LDL 콜레스테롤이 1 단위 증가할 때 오즈가 1 보다 작기 때문에 이 경우에서는 'LDL 콜레스테롤' 변수의 증가는 관상동맥질환 발병률을 유의미하게 감소시킨다고 볼 수 있습니다. 앞서 설명변수를 분석한 것 과는 상반된 결과가 나옵니다.

5. 예측 모형 적합

예측 모형을 세우기 위해 우선 아래의 box-plot 을 통해 이상 값들을 제거합니다. box-plot 은 왼쪽 위에서부터 age, TCHOL, hdl, LDL, hemoglobin, WBC, neutrophilbox, lymphocyte 입니다. 따라서 위의 box-plot 을통해 이상관측치를 제거함으로 데이터를 정제합니다. 그 후에 stepAIC 를 통해 AIC 가 가장 낮은 모형을 적합 시킨 후 위에서 적합 시킨 회귀식의 AIC 와 비교합니다.



```
Call:
glm(formula = CADGROUP ~ age + Sex + smoking + LDL + WBC + lymphocyte +
    DM, family = binomial(link = "logit"), data = dataa, trace = TRUE)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9696  -0.8144  -0.5071   0.9218   2.5667

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.437e+00  9.166e-01  -2.659  0.00785 **
age          3.765e-02  8.923e-03   4.220 2.44e-05 ***
Sexman      -1.804e+00  2.845e-01  -6.339 2.31e-10 ***
smokingTRUE  2.097e+00  2.934e-01   7.148 8.84e-13 ***
LDL         -5.568e-03  3.198e-03  -1.741  0.08163 .
WBC          8.917e-05  5.006e-05   1.781  0.07489 .
lymphocyte  -2.951e-02  1.102e-02  -2.677  0.00744 **
DMTRUE       5.868e-01  2.111e-01   2.780  0.00543 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 823.16  on 664  degrees of freedom
Residual deviance: 679.63  on 657  degrees of freedom
AIC: 695.63

Number of Fisher Scoring iterations: 5
```

```
Call:
glm(formula = CADGROUP ~ age + log(neutrophil/lymphocyte) + WBC +
    smoking + LDL + LDL:Sex + LDL:DM, family = binomial(link = "logit"),
    data = dataa, trace = TRUE)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0615  -0.8208  -0.4989   0.8705   2.8985

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)               -4.299e+00  7.624e-01  -5.639 1.71e-08 ***
age                        4.271e-02  9.036e-03   4.726 2.28e-06 ***
log(neutrophil/lymphocyte) 4.475e-01  1.989e-01   2.250 0.024444 *
WBC                        8.418e-05  5.112e-05   1.647 0.099613 .
smokingTRUE                2.031e+00  2.806e-01   7.238 4.56e-13 ***
LDL                       -1.116e-03  3.365e-03  -0.332 0.740201
LDL:Sexman                -2.150e-02  3.191e-03  -6.738 1.61e-11 ***
LDL:DMTRUE                 8.359e-03  2.471e-03   3.384 0.000716 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 823.16  on 664  degrees of freedom
Residual deviance: 673.89  on 657  degrees of freedom
AIC: 689.89

Number of Fisher Scoring iterations: 5
```

AIC 를 비교했을 때 기존의 모형의 AIC 가 더 낮으니 기존 모형을 사용합니다.

예측도 검사를 위해 자료의 1/9 를 valid 자료로 랜덤하게 추출하고, 나머지 8/9 를 train 자료로 사용합니다. 그리고 train 자료에서 1/2 의 자료를 한번 더 추출해 그 데이터를 사용해 예측 모형을 세운 뒤 가장 큰 예측력을 가지는 모형을 예측모형으로 선택하도록 합니다.

(Intercept)	age	log(neutrophil/lymphocyte)	WBC
-3.180947e+00	2.692292e-02	4.812328e-01	2.908111e-05
smokingTRUE	LDL	LDL:Sexman	LDL:DMTRUE
2.242741e+00	3.664049e-03	-2.635925e-02	8.107566e-03

따라서 예측력이 0.7918919 로 가장 큰 모형을 선택하면 다음 모형이 적합합니다.

$$\hat{Y} = -3.180947 + 0.0262292\,age + 0.4812328\log\left(\frac{neutrophil}{lymphocyte}\right) + 0.00002908111\,WBC + 2.242741\,smoking + 0.003664049\,LDL - 0.0263592\,LDL \times Sex + 0.008107566\,LDL \times DM$$

(단, 남성일 때 Sex 가 1, 여성일 때 Sex 가 0, 흡연자일 때 smoking 이 1, 비흡연자일 때 smoking 이 0, 당뇨환자일 때 DM 이 1, 정상일 때 DM 이 0, 파이는 관상동맥질환 발병률 의미)

인용 자료

http:///jkna.org/upload/pdf/201003005.pdf

https://dspace.inha.ac.kr/bitstream/10505/19169/1/20105.pdf

https://terms.naver.com/entry.nhn?docId=3535926&cid=58572&categoryId=58572.

https://www.huffingtonpost.kr/2016/03/18/story_n_9493302.html.