[회귀분석2 최종보고서]

- 로지스틱 회귀모형을 이용한 심장병 발생확률 분석 -

서울시립대학교

회귀분석2 (김규성 교수님)

2016580009 통계학과 김태현

차례

1. 서론

1.1 연구 목적

심장병은 심장을 구성하고 있는 구조나 기능에 이상이 생긴 것으로, 임신 중 심장이 생기는 과정에서 문제가 발생한 경우인 선천성 심장병과 정상 심장으로 태어나 동맥경화증, 고혈압, 잘못된 식생활습관 등에 의해 발생하는 후천성 심장병이 있다. 심장병의 종류에는 협심증, 심근경색, 동맥경화증, 심장판막증, 심부전 등이 있다. (본 연구에서는 종류를 구분하지 않고 심장병 진단 여부를 분석한다.)

심장병은 전 세계 인구의 가장 큰 사망원인 중 하나이고 미국에서만 매년 약 610,000명이 심장병으로 사망한다. 따라서 심장병 발생의 예측은 의료 산업의 데이터 분석 분야에서 중요한 부분이지만 당뇨병, 혈압, 콜레스테롤, 통증의 정도, 맥박 등 많은 요인이 복잡하게 얽혀 있어서 단순 비교로는 심장질환을 분석하기 쉽지 않

다. 따라서 통계 기법 등을 이용한 고차원적인 접근 방식이 필요하다. 본 연구에서는 여러 가지 요인들을 바탕으로 심장병 진단 여부를 분석하여 심장병 진단에 중요한 요인이 무엇인지 판단하고 심장병을 진단 여부를 확률적으로 알 수 있게 한다.

1.2 문헌 연구

Regression methods for analyzing the risk factors for a life style disease among the young population of India, B. Ismail and Manjula Anil라는 연구에서 심장병의 일종인 관상동맥질환(Coronary artery disease)을 반응변수로 하고 나이(Age), 성별(Sex), BMI(체질량 지수), Cholesterol(콜레스테롤 수치), HDL(고밀도 리포 단백질; 좋은 콜레스테롤), LDL(저밀도 리포 단백질; 나쁜 콜레스테롤) 그리고 혈압(Blood Pressure)을 설명변수로 하여 로지스틱 회귀분석을 실시하였다.

Relation of Helicobacter pylori infection and coronary heart disease. M. A. Mendall, P. M. Goggin, N. Molineaux, J. Levy, T. Toosy, D. Strachan, A. J. Camm, T. C. Northfield라는 연구에서는 헬리코박터 파일로리균과 관상동맥질환의 관계에 대해서 로지스틱 회귀분석을 사용했다.

Comparing performances of logistic regression, decision trees, and neural networks for classifying heart disease patients, IEEE, A. Khemphila라는 연구에서는 심장병 환자의 분류를 위해 로지스틱 회귀분석(Logistic Regression), 의사결정 나무(Decision Tree), 딥러닝(Neural Networks) 등 모델들의 성능을 비교하였다.

1.3 데이터 설명

데이터 이름은 Cleveland Heart Disease Dataset 이다. 원출처는 UCI Repository이고 주요 변수 8개를 분석 데이터로 사용할 것이다.

데이터의 행 개수는 1,025개로 총 1,025명의 환자의 데이터가 사용되었다. 열 개수는 8개이며 저장형식은 csv이다.

변수 1은 Age이다. 개인의 나이를 나타내는 연속형 변수이다.

변수 2는 Sex이다. 남성이면 1을, 여성이면 0을 가지는 이산형 변수이다.

변수 3은 cp(chest-pain type)이다. 가슴에 느껴지는 통증의 유형에 따라서 4단계로 구분되는 범주형 변수이다. 전형적인 협심증(typical angina)은 0, 전형적이지 않은 협심증(atypical angina)는 1, 협심증이 아닌 통증(Non-anginal pain)은 2 그리고 점근적인 통증(asymptotic)은 3의 값을 가진다.

변수 4는 trestbps(Resting Blood Pressure)이다. 휴식상태에서의 혈압을 나타내는 연속형 변수이다. 단위는 mmHg이다.

변수 5는 chol(Serum Cholesterol)이다. 혈액 속에 심장에 "나쁜" 콜레스테롤인 low-density lipoprotein cholesterol이 얼마나 들어있는지를 나타내는 연속형 변수이다. 단위는 mg/dl이다.

변수 6은 fbs(Fasting Blood Sugar)이다. 혈당이 120을 기준으로 그보다 높으면 1, 낮으면 0의 값을 갖는 이산형 변수이다. 단위는 mg/dl이다.

변수 7은 thalach(Maximum Heart Rate Achived)이다. 최대 심박수를 나타내는 연속형 변수이다.

변수 8은 target이다. 심장병이 있다고 진단되었으면 1, 그렇지 않으면 값 0를 가지는 이산형 변수이다. 본 연구에서는 이 변수에 대한 로짓을 반응변수로 하여 심장병 진단받을 확률을 추정할 것이다.

1.4 분석 방법

본 연구는 R과 SAS를 사용하여 분석을 진행한다.

연구에서는 반응 변수(target)가 심장병 진단 여부에 따라 1 또는 0의 값을 가지는 이산형 변수이다. 로지스틱 회귀분석은 반응 변수가 이산형일 때 적합한 분석 방법이라고 알려져 있다.

일반적인 선형 회귀모형과 비교했을 때 로지스틱 회귀모형을 사용하는 가장 큰 이유는 다음과 같다. 선형 회귀모형의 경우 결괏값이 0과 1 사이를 벗어나 버리는 문제가 발생한다. 로지스틱 회귀모형은 로짓(log odds ratio)를 이용해 결과값을 0과 1 사이로 제한한다. 또한 일반 선형모형을 적합할 경우 0 또는 1의 값을 갖는 y값에 대응되는 오차가 더 이상 정규성을 갖지 못한다.

따라서 로지스틱 회귀분석을 이용하여 어떠한 사람이 심장병을 갖고 있을 확률을 분석하고 각 요인이 심장병 보유에 미치는 영향을 추정한다.

다음과 같은 로지스틱 회귀모형의 회귀식을 사용하여 각 변수의 회귀계수와 확률값을 추정한다.

$$x = (x_1, x_2, \cdots, x_{13})'$$

$$y = (y)'$$

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = x'\beta = \beta_0 + \beta_1 x_1 + \cdots + \beta_7 x_7, \ \ \pi(x) = P(y=1|x)$$

$$\pi(x) = \frac{1}{1+e^{-x'\beta}}$$

모형의 적합성은 우도비 검정을 통해 검정하며 검정통계량은 다음과 같다.

$$-2\ln(\frac{L(R)}{L(F)}) \sim \chi^2(\ )$$

여기서 $L(R)$은 제약모형 하에서의 우도비이고 $L(F)$은 완전모형 하에서의 우도비이다. 또한, 검정통계량은 대표본이론에 따라 카이제곱분포를 따른다. 검정 통계량의 값이 클수록 추정된 로지스틱 회귀분석 모형이 적합하다고 나타난다.

1.5 결과 활용 및 기대 효과

분석 결과로 추정된 로지스틱 회귀모형이 나오게 되면 다른 복잡하고 고비용의 검사들에 비해 비교적 간단한 검사를 통해 얻을 수 있는 변수들인 연령, 성별, 혈압, 혈당 등을 바탕으로 심장병 진단 여부를 예측할 수 있게 된다. 또한 각각의 설명변수들이 심장병에 미치는 영향력을 판단할 수 있다.

가정에서도 간단한 검사들로 심장병을 진단하는 데 도움을 준다. 이를 통해서 심장병 검사를 위해 의료시설에 방문하기 전에 심장병을 가질 위험도를 미리 대략적으로 판단해 볼 수 있게 된다.

2. 본론

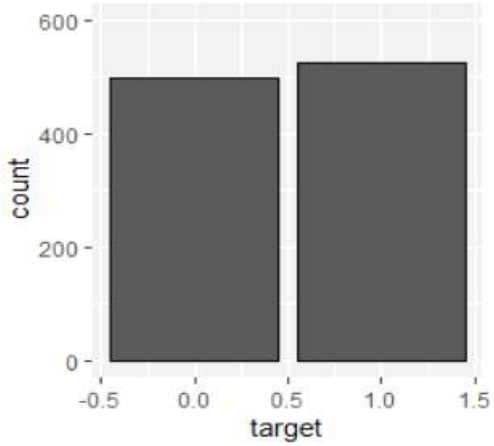
2.1 분석 방법 소개

우선 변수들의 특성에 대해 기초통계 분석을 진행한다. 그 후 개별 설명변수들에 대한 독립성 검정을 시행하여 각 설명변수와 반응변수 target의 관계를 확인한다.

모든 변수를 포함한 로지스틱 회귀분석을 진행하여 모형의 적합성과 회귀계수의 유의성을 확인한다. 그 결과를 기반으로 stepwise 단계선택법을 통해 어떤 변수를 선택할 것인지 결정한다. 그 후 우도비 검정, Cross validation 등의 모형 진단, 평가 방법으로 추정된 모형의 타당성을 확인한다.

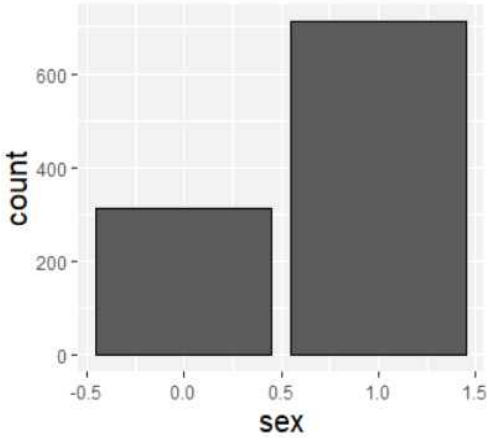
2.2 데이터 분석 및 결과 설명

변수들에 대한 기초통계 분석은 다음과 같다.



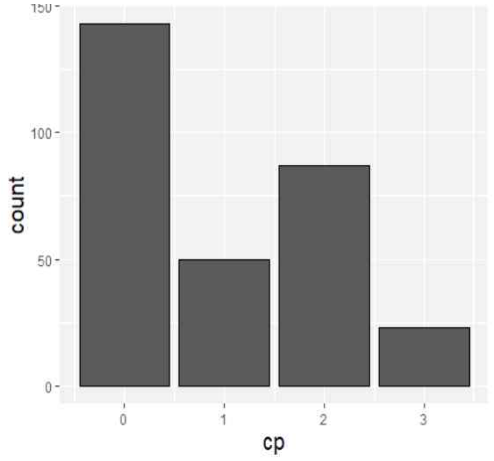
0(심장병 없음)	1(심장병 진단)
498	526

분석 시 반응변수 y에 해당하는 target 변수이다. 이산형 변수이며 심장병을 진단받지 않은 환자는 498명, 심장병이 있다고 진단된 환자는 526명이다.



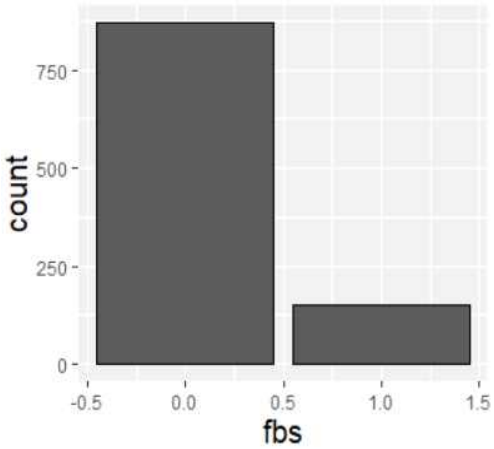
0(여성)	1(남성)
312	713

sex는 이산형 변수이며 여성은 312명 남성은 713명으로 남성이 여성보다 약 2배 정도 많은 표본을 갖는다.



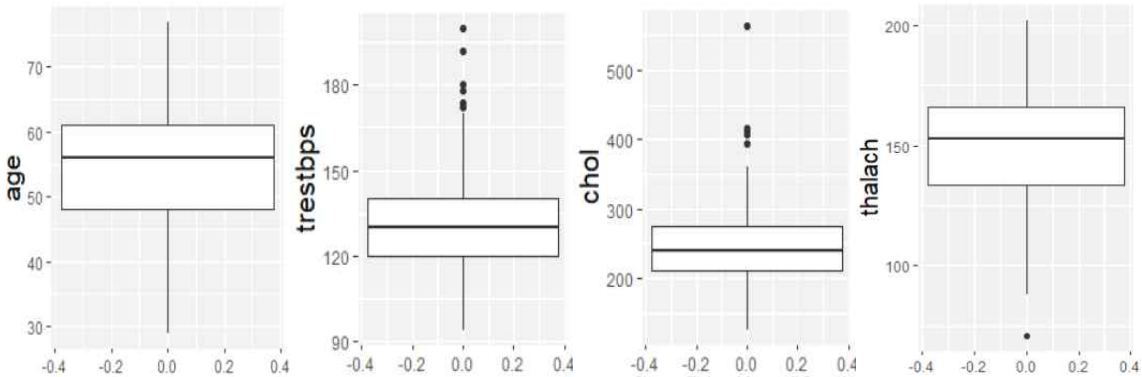
0 (전형적인 협심증)	1 (전형적이지 않은 협심증)	2 (협심증이 아닌 통증)	3 (점근적인 통증)
497	167	284	77

cp는 이산형 변수이며 표본중에 497명은 전형적인 협십증 정도의 통증을 호소하는 환자들이다. 167명은 전형적이지 않은 협십증 증상을, 284명은 협심증이 아닌 통증 그리고 77명은 점근적인 통증을 호소한다. 4개의 값을 가지는 범주형 변수이으모 회귀분석 시에 범주형 변수로 지정을 해주어 분석을 진행해야 한다.

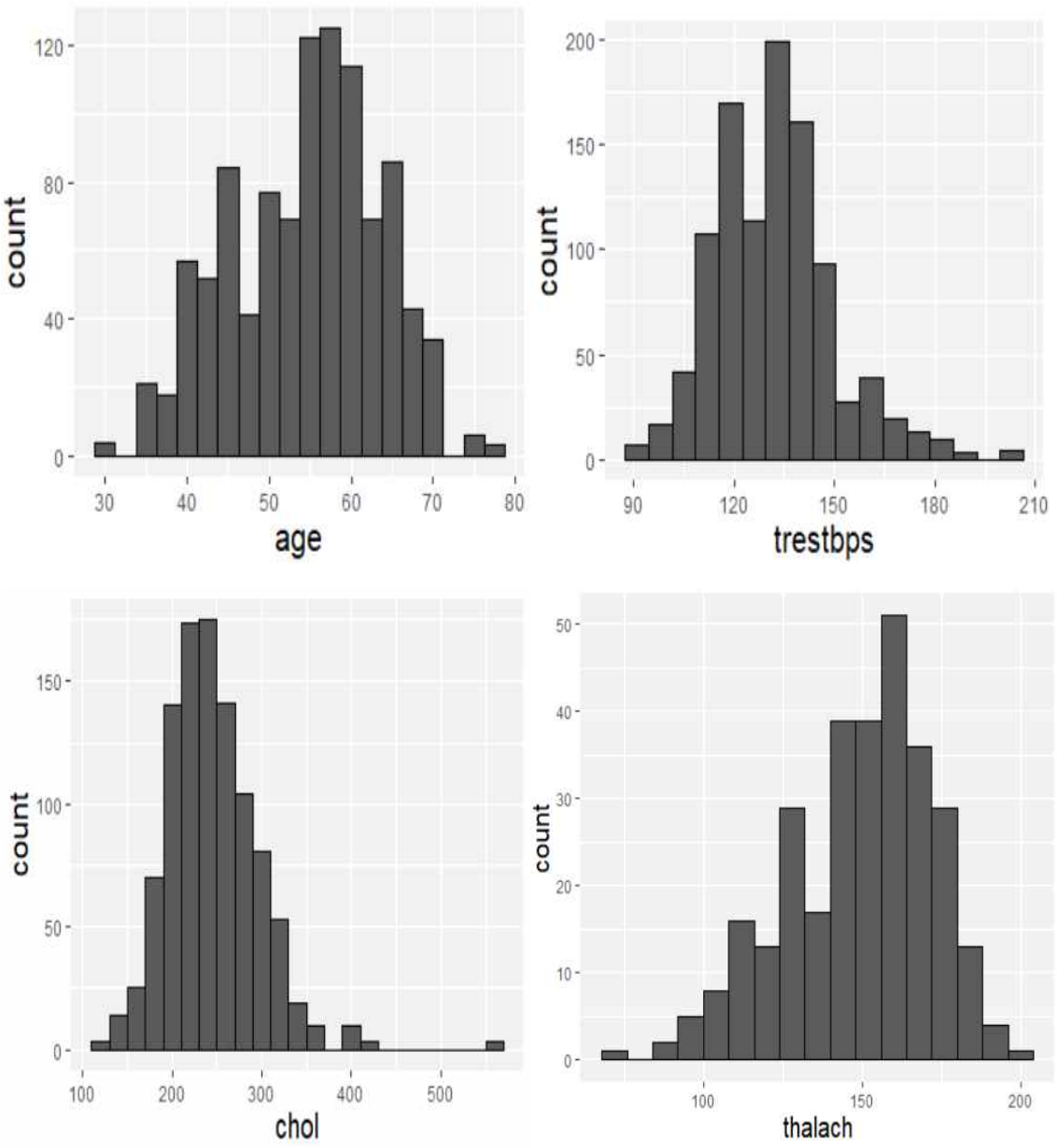


0 (혈당 120이하)	1 (혈당 120이상)
872	153

fbs는 이산형 변수이며 혈당을 기준으로 120mg/dl보다 높은 환자의 수는 153명이고 그보다 낮은 환자의 수는 372명이다. 그 수치가 120 이하인 환자가 120 이상인 환자가 약 2.5배 많다.



변수	age	trestbps	chol	thalach
평균	54.37	131.62	246.26	149.65
표준편차	9.08	17.54	17.54	22.91
1분위수	47.5	120	211	133.5
3분위수	61	140	274.5	166
최소값	29	94	126	71
최대값	77	200	564	202



age는 연속형 변수이며 평균은 54.37세이다. 심장병을 검사하는 연령이 상당이 높은 편임을 알 수 있다.

trestbps는 연속형 변수이며 평균은 131.62로 환자의 대부분은 혈압이 평균 근처에

밀집되어 있다.

chol는 연속형 변수이며 변수의 평균은 246.26이고 정규분포와 유사한 형태를 띄고 있다. 최대값인 564를 갖는 환자 한 명은 이상치로 판단되며 제거해주고 분석을 진행하겠다.

thalach은 연속형 변수이며 변수의 평균은 149.65이다. 네 개의 연속형 변수 중 산포(표준편차)가 가장 크다.

변수	N	평균	표준편차	최솟값	최댓값
age	303	54.3663366	9.0821010	29.0000000	77.0000000
sex	303	0.6831683	0.4660108	0	1.0000000
cp	303	0.9669967	1.0320525	0	3.0000000
trestbps	303	131.6237624	17.5381428	94.0000000	200.0000000
chol	303	246.2640264	51.8307510	126.0000000	564.0000000
fbs	303	0.1485149	0.3561979	0	1.0000000
thalach	303	149.6468647	22.9051611	71.0000000	202.0000000
target	303	0.5445545	0.4988348	0	1.0000000

각 설명 변수(age, sex, cp, trestbps, chol, fbs)별로 반응변수(target)와 $H_0 : \beta = 0$을 귀무가설로 하는 카이제곱 독립성 검정을 시행한다.

age vs target

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	55.1747	1	<.0001
Score	53.9040	1	<.0001
Wald	51.5538	1	<.0001

우도비 검정통계량이 55.1474이고 p-value가 0.0001보다 작다. 귀무가설을 기각하며 나이는 심장병 진단과 관계가 있다.

sex vs target

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	82.3927	1	<.0001
Score	80.0737	1	<.0001
Wald	75.8193	1	<.0001

우도비 검정통계량이 82.3927이고 p-value가 0.0001보다 작다. 귀무가설을 기각하며 성별은 심장병 진단과 관계가 있다.

cp vs target

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	85.9412	3	<.0001
Score	81.6864	3	<.0001
Wald	72.5704	3	<.0001

우도비 검정통계량이 85.9412이고 p-value가 0.0001보다 작다. 귀무가설을 기각하며 통증 유형은 심장병 진단과 관계가 있다.

trestbps vs target

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	19.9586	1	<.0001
Score	19.7390	1	<.0001
Wald	19.2517	1	<.0001

우도비 검정통계량이 19.9586이고 p-value가 0.0001보다 작다. 귀무가설을 기각하며 혈압은 심장병 진단과 관계가 있다.

chol vs target

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	10.3409	1	0.0013
Score	10.2429	1	0.0014
Wald	10.0380	1	0.0015

우도비 검정통계량이 10.3409이고 p-value가 0.0013이다. 귀무가설을 기각하며 콜레스테롤 수치는 심장병 진단과 관계가 있다.

fbs vs target

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	1.7367	1	0.1876
Score	1.7368	1	0.1875
Wald	1.7299	1	0.1884

우도비 검정통계량이 1.7367이고 p-value가 0.1876이다. 귀무가설을 채택하며 혈당이 120을 넘는 여부는 심장병 진단과 관계가 없다고 볼 수 있다.

thalach vs target

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	58.3811	1	<.0001
Score	53.8932	1	<.0001
Wald	45.2802	1	<.0001

우도비 검정통계량이 58.3811이고 p-value가 0.0001보다 작다. 귀무가설을 기각하며 최대 심박수 수치는 심장병 진단과 관계가 있다.

각 설명변수들의 독립성 검정 결과 fbs 변수는 심장병 진단에 영향을 주지 않으며 fbs 변수를 제외한 나머지 설명변수들은 관계가 있다는 결론을 내렸다. 우도비 검정통계량을 바탕으로 결론을 내렸으며 score 검정통계량, wald 검정통계량을 고려해도 같은 결과를 도출한다.

다음과 같은 모든 설명변수를 포함하는 모형을 적합해 위 결과를 확인해 본다.

$\pi(x)$: 심장병이라고 진단받을 확률

$$\ln\left(\frac{\widehat{\pi(x)}}{1-\widehat{\pi(x)}}\right) = \widehat{\beta_{age}} \times age + \widehat{\beta_{sex}} \times sex + \widehat{\beta_{cp0}} \times cp0 + \widehat{\beta_{cp1}} \times cp1 + \widehat{\beta_{cp2}} \times cp2 + \widehat{\beta_{trestbps}} \times trestbps + \widehat{\beta_{chol}} \times chol + \widehat{\beta_{fbs}} \times fbs + \widehat{\beta_{thalach}} \times thalach$$

적합 결과는 다음과 같다.

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-2.4927	2.1026	1.4056	0.2358
age		1	0.0227	0.0202	1.2592	0.2618
sex		1	2.0525	0.3856	28.3351	<.0001
cp	0	1	1.6160	0.2560	39.8342	<.0001
cp	1	1	-0.5284	0.3550	2.2158	0.1366
cp	2	1	-0.5124	0.2823	3.2957	0.0695
trestbps		1	0.0230	0.00932	6.0793	0.0137
chol		1	0.00658	0.00310	4.4973	0.0339
fbs		1	0.0998	0.4284	0.0542	0.8159
thalach		1	-0.0373	0.00875	18.1468	<.0001

$\widehat{\beta_0} = -2.4927\,(p-value = 0.2358)$
$\hat{\beta}_{age} = 0.0227\,(p-value = 0.2618)$
$\hat{\beta}_{sex} = 2.0525\,(p-value < 0.0001)$
$\hat{\beta}_{cp0} = 1.6160\,(p-value < 0.0001)$
$\hat{\beta}_{cp1} = -0.5284\,(p-value = 0.1366)$
$\hat{\beta}_{cp2} = -0.5124\,(p-value = 0.0695)$
$\hat{\beta}_{trestbps} = 0.0230\,(p-value = 0.0137)$
$\hat{\beta}_{chol} = 0.00658\,(p-value = 0.0339)$
$\hat{\beta}_{fbs} = 0.0998\,(p-value = 0.8159)$
$\hat{\beta}_{chol} = -0.0373\,(p-value < 0.0001)$

위 결과에 따르면 y 절편, age, fbs 변수에 대한 유의성 검정의 p-value가 각각 0.2358, 0.2618, 0.8159이다. 따라서 귀무가설 $H_o : \beta = 0$을 기각하지 못하게 되며 이 변수들은 유의하지 않다는 결론을 내릴 수 있다. 그 외 나머지 변수들은 귀무가설을 기각하며 심장병 진단 확률을 예측하는 데에 유의미한 변수라고 볼 수 있다.

age 변수는 개별 유의성 검정에서는 유의하지만 모든 변수를 넣은 모델에서는 유의하지 않은 결과가 나타났다. 그 이유를 다른 변수와의 상관관계로 보고 변수들의 상관계수를 확인하였다.

피어슨 상관 계수, N = 303 H0: Rho=0 가정하에서 Prob > |r|								
	age	sex	cp	trestbps	chol	fbs	thalach	target
age	1.00000	-0.09845 0.0871	-0.06865 0.2335	0.27935 <.0001	0.21368 0.0002	0.12131 0.0348	-0.39852 <.0001	-0.22544 <.0001

age 변수는 trestbps 변수와 양의 상관관계(p-value <0.0001), thalach 변수와 음의 상관계(p-value <0.0001)이 있음이 확인되었다. 즉, age 변수는 target에 영향을 주지만 상관관계가 있으며 모델을 더 잘 설명하는 다른 변수들을 선택하고 변수선택 단계에서 제거하겠다.

위 결과를 한 번 더 확인하기 위해 Stepwise 단계적 변수 선택법을 통한 변수 선택을 진행하는 상수항이 없는 로지스틱 회귀모형을 적합해 본다. stepwise 변수 선택법은 모든 변수가 포함된 모델에서 출발하여 기준 통계치에 가장 도움이 되지 않는 변수를 삭제하거나, 모델에서 빠져 있는 변수 중에서 기준 통계치를 가장 개선시키는 변수를 추가하는 과정을 반복하며 최적의 모형을 찾아가는 방법이다.

Summary of Stepwise Selection							
Step	Effect Entered	Effect Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
1	cp		3	1	66.9430		<.0001
2	thalach		1	2	24.2854		<.0001
3	sex		1	3	30.5228		<.0001
4	trestbps		1	4	20.6694		<.0001
5	chol		1	5	4.8077		0.0283

그 결과 age, fps 변수가 stepwise 방법을 통해서 제거되었으며 최종모형은 다음과 같다.

$\pi(x)$: 심장병이라고 진단받을 확률

$$\ln\left(\frac{\widehat{\pi(x)}}{1-\widehat{\pi(x)}}\right) = \widehat{\beta_{sex}} \times sex + \widehat{\beta_{cp0}} \times cp0 + \widehat{\beta_{cp1}} \times cp1 + \widehat{\beta_{cp2}} \times cp2 + \widehat{\beta_{trestbps}} \times trestbps + \widehat{\beta_{chol}} \times chol + \widehat{\beta_{thalach}} \times thalach$$

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
sex		1	1.9483	0.3690	27.8796	<.0001
cp	0	1	1.5511	0.2458	39.8282	<.0001
cp	1	1	-0.5283	0.3511	2.2639	0.1324
cp	2	1	-0.5197	0.2768	3.5248	0.0605
trestbps		1	0.0223	0.00700	10.1983	0.0014
chol		1	0.00637	0.00291	4.7825	0.0288
thalach		1	-0.0438	0.00709	38.0701	<.0001

$$\ln\left(\frac{\widehat{\pi(x)}}{1-\widehat{\pi(x)}}\right) = 1.9483 \times sex + 1.5511 \times cp0 - 0.5283 \times cp1 - 0.5197 \times cp2 + 0.0223 \times trestbps + 0.00637 \times chol - 0.0438 \times thalach$$

성별(sex)이 여성(0)일 때보다 남성(1)일 때 심장병이라고 진단받을 로짓이 1.9483 만큼 높다.

통증 유형(cp)이 전형적인 협심증(0)일 때 점근적인 통증(3)에 비해서 심장병이라고 진단받을 로짓이 1.5511 높다. 전형적이지 않은 협심증(1)일 때는 (3)보다 –0.5283만큼 낮고 협심증이 아닌 통증(2)일 때는 (3)의 경우보다 0.5197만큼 낮다.

혈압(trestbps)은 1mmHg 증가할 때마다 심장병이라고 진단받을 로짓이 0.0223 높아진다.

콜레스테롤 수치(chol)이 1mg/dl 증가할 때마다 심장병이라고 진단받을 로짓은 0.00637 증가한다.

최대 심박수(thalach)가 1 증가할 때마다 심장병이라고 진단받을 로짓은 0.0438 감소한다.

예를 들어, 성별이 여성, 전형적인 협심증 정도의 통증을 가지며, 혈압이 131이고

콜레스테롤 수치가 246, 최대 심박수가 149인 환자가 있을 때 그 환자가 심장병이라고 진단받을 로짓은 1.9483(1) + 1.5511(1) - 0.5283(0) - 0.5197(0) + 0.0223(131) + 0.00637(246) - 0.0438(149) = 1.46152이다. ($\ln(\frac{\widehat{\pi(x)}}{1-\widehat{\pi(x)}}) = 1.46152$)

즉, $\widehat{\pi(x)} \simeq 0.812$이며 이 여성이 심장병이 진단받을 확률은 81.2%로 추정된다.

2.3 분석의 타당성 설명

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	153.0722	7	<.0001
Score	124.6712	7	<.0001
Wald	79.6576	7	<.0001

귀무가설이 $H_0 : \beta_1 = \beta_2 = \cdots = \beta_7 = 0$인 우도비 검정을 진행하였다. 그 결과 우도비 통계량은 153.0722, p-value <0.0001로 귀무가설을 기각하게 된다. 즉, 심장병 진단은 7개의 변수 중 적어도 한 변수에 의존한다.

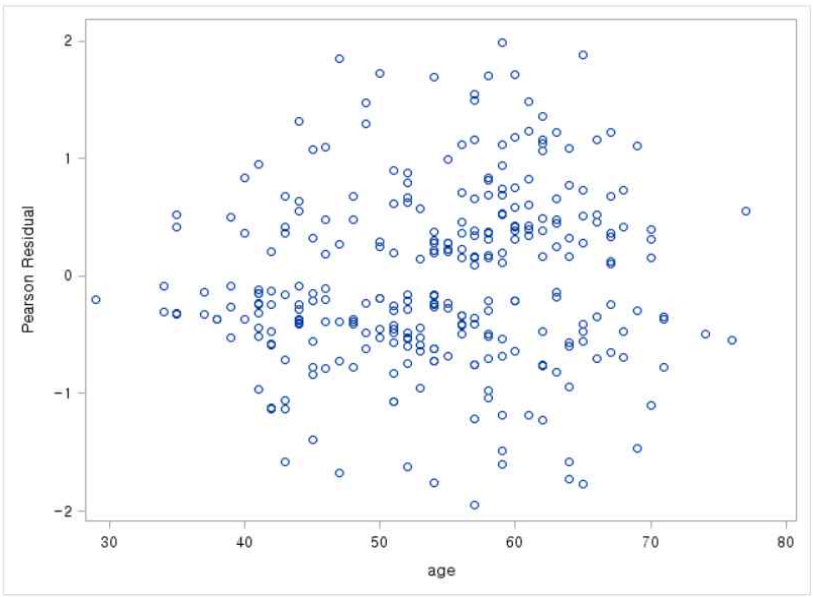
Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
sex	1	27.8796	<.0001
cp	3	44.9075	<.0001
trestbps	1	10.1983	0.0014
chol	1	4.7825	0.0288
thalach	1	38.0701	<.0001

각 회귀계수에 대해서 유의성 검정을 진행하였다. sex 변수의 Wald 검정통계량은 27.8796(p-value < 0.0001), cp 변수의 Wald 검정통계량은 44.9075(p-value < 0.0001), trestbps 변수의 Wald 검정통계량은 10.1983(p-value = 0.0014), chol 변

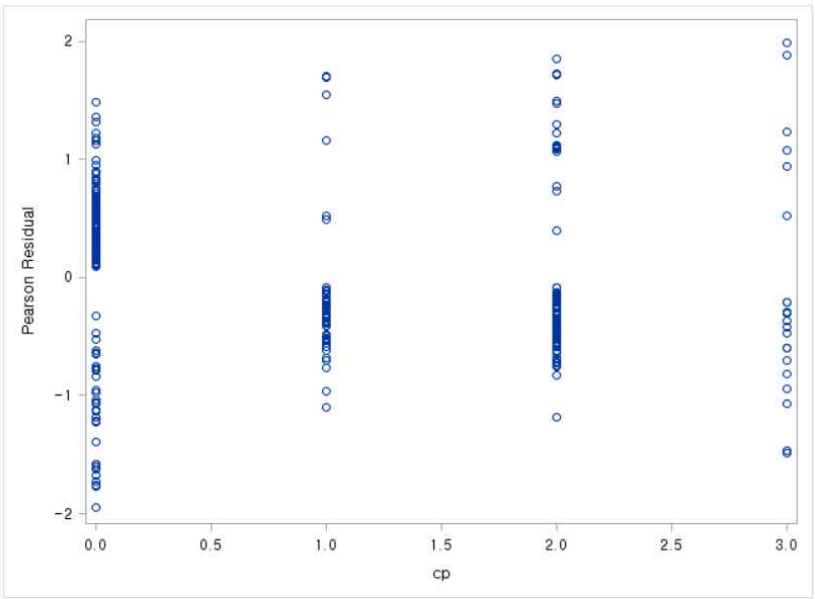
수의 Wald 검정통계량은 4.7825(p-value = 0.0288) 그리고 thalach 변수의 Wald 검정통계량은 38.0701(p-value < 0.0001)이다.

따라서, 최종모형에 사용된 모든 변수는 귀무가설 $H_0 : \beta = 0$을 기각한다. 즉, 나이, 통증 유형, 혈압, 콜레스테롤, 혈당, 최대 심박수는 심장병 진단에 유의한 변수들이다.

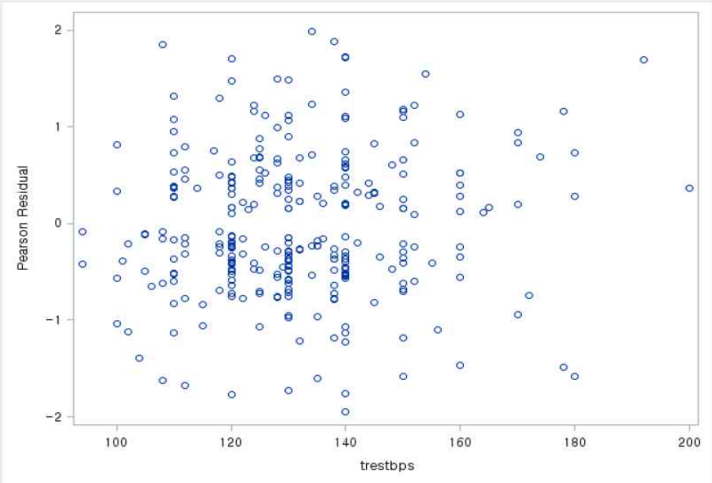
설명변수와의 잔차 그림을 통해서 모형이 잘 적합됐는지 판단한다.



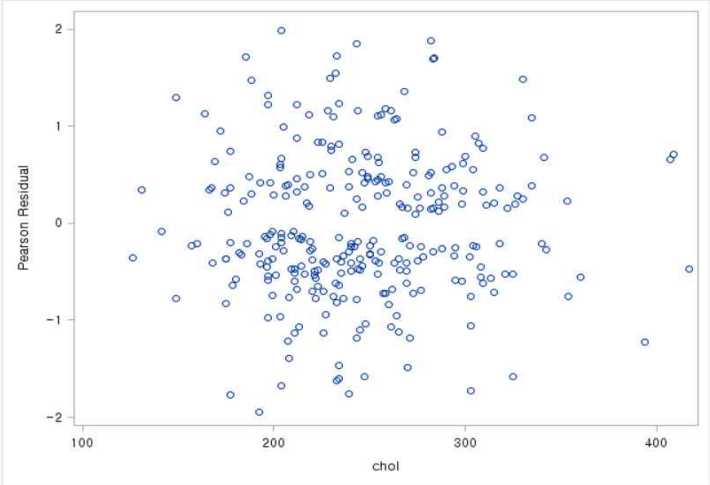
특별한 패턴이 없다.



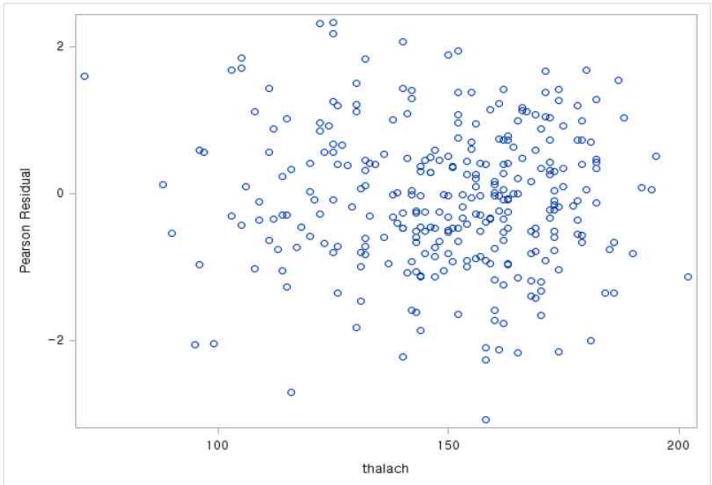
특별한 패턴이 없다.



특별한 패턴이 없다.



특별한 패턴이 없다.



특별한 패턴이 없다.

잔차 그림에서 특별한 패턴을 발견하지 못했다. 따라서 모형은 적합하다.

Partition for the Hosmer and Lemeshow Test					
		target = 0		target = 1	
Group	Total	Observed	Expected	Observed	Expected
1	30	1	0.88	29	29.12
2	30	3	2.57	27	27.43
3	30	3	4.54	27	25.46
4	30	7	6.85	23	23.15
5	30	7	10.01	23	19.99
6	30	18	13.87	12	16.13
7	30	20	18.82	10	11.18
8	30	20	22.93	10	7.07
9	30	29	26.28	1	3.72
10	33	30	31.64	3	1.36

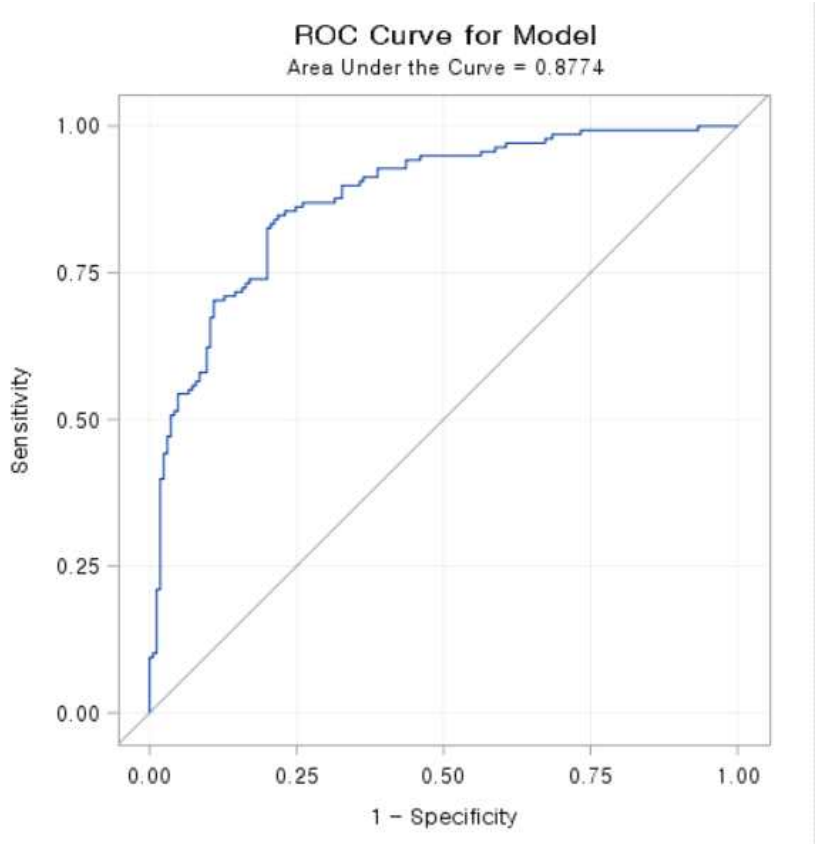
Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
10.4916	8	0.2322

Homer - Lemeshow 적합도 검정은 로지스틱 회귀모형에서의 적합도 검정으로 그룹으로 나누어 관측 빈도와 기대 빈도를 비교하여 모형에 데이터가 얼마나 잘 적합하는지 평가한다. 귀무가설은 관측된 비율과 기대되는 비율이 동일하다는 것으로 둔다.

2.2에서의 최종모형에 대한 적합도 검정을 시행했을 때 카이제곱 통계량이 10.4916이고 p-value가 0.2322로 귀무가설을 기각하지 못한다는 검정 결과가 나타났다. 즉, 귀무가설을 채택하게 되며 이는 최종모형이 적합하다는 것을 의미한다.

Classification Table									
	Correct		Incorrect		Percentages				
Prob Level	Event	Non-Event	Event	Non-Event	Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
0.500	72	92	21	27	77.4	72.7	81.4	22.6	22.7

모델의 정확도를 확인하기 위해서 원 데이터를 train 데이터와 test 데이터를 7:3으로 나누어서 cross validation을 진행하였다. 그 결과 정확도는 77.4%가 나왔으며 꽤 높은 수준의 정확도를 보여준다.



ROC 곡선은 민감도와 특이도로 그려지는 곡선을 의미하며 모델의 정확도의 지표가 된다. 곡선의 밑면적 즉, AUC(Area Under the Curve)가 클수록 모델의 성능이 좋음을 의미한다. 최종모형의 AUC는 0.8774로 높은 수준의 성능을 보여준다.

3. 결론

3.1 분석 결과 요약

나이, 성별, 통증 유형, 혈압, 콜레스테롤 수치, 혈당 수치가 120보다 높은지 여부, 최대 심박수를 이용하여 심장병 진단 여부 확률을 추정하는 분석을 진행하였다.

그 결과 나이(age)는 심장병 진단에 영향을 주는 변수였지만 다른 변수들과의 상관관계를 고려하여 모형에서 제외되었고 혈당 수치가 120보다 높은지 여부(fbs)는 유의하지 않은 변수로 판단되어 모형에 포함되지 못했다.

최종모형은 성별, 통증 유형, 혈압, 콜레스테롤 수치, 최대 심박수를 설명변수로 하는 로지스틱 회귀모형이며 적합 후 모형 진단 결과 타당한 모형으로 판단되었다. 최종 모형은 다음과 같다.

$$\ln\left(\frac{\widehat{\pi(x)}}{1-\widehat{\pi(x)}}\right) = 1.9483 \times sex + 1.5511 \times cp0 - 0.5283 \times cp1 - 0.5197 \times cp2 + 0.0223 \times trestbps + 0.00637 \times chol - 0.0438 \times thalach$$

위 결과를 요약하면 여성이 심장병 진단받을 확률이 더 높고, 전형적인 협심증 유형의 통증이 나타날 때 높다. 또한 혈압, 콜레스테롤 수치는 높을수록 심장병 진단 확률이 높고 반대로 최대 심박수는 낮을수록 그 확률이 높았다.

3.2 분석의 장점 및 한계점 설명

로지스틱 회귀분석의 장점은 범주가 2개인 이산형 변수에 대해서 확률적으로 표현할 수 있다는 것이다. 그로 인해 원래는 값이 1 또는 0을 갖는 심장병 진단을 좀 더 구체적인 수치로 확인할 수 있게 되었다.

또한 통증 유형, 혈압 등과 같은 수치는 의료시설에서의 좀 더 복잡하고 고비용인 검사들에 비해 비교적 쉽게 얻을 수 있는 정보들이다. 따라서 그렇게 많은 시간과 비용을 들이지 않더라도 심장병 진단에 대한 정보를 얻을 수 있다.

데이터가 미국인에 한정되어 있다는 점 때문에 많은 인종에 대입하기 어려울 수 있고 혈당 변수의 경우 120을 기준으로 높으면 1, 낮으면 0을 갖는 이산형 변수가 아닌 구체적으로 연속형 변수를 사용했다면 더 개선된 모형을 얻을 수 있을 것 같다. 여러 인종의 구체적인 변수를 사용하지 못한 점이 아쉬움으로 남는다.

이진 변수에 대한 다른 모형에는 probit 모형, complementary log-log model 등이 있다. 이 모형들도 로지스틱 회귀모형과 마찬가지로 π를 설명변수 x로 설명하고자 하는 모형들이다. 더 다양한 모형을 시도해서 좀 더 정확도가 높은 모형을 채택한다면 더 나은 결과를 얻을 수 있을 것이다.

C. 참고문헌

1. 서울시립대학교 통계학과 김규성 교수님 회귀분석2 강의자료
2. 유한양행
https://www.yuhan.co.kr/Mobile/Introduce/Health/index.asp?mode=view&Cateid=290&IDX=2634&p=136&sm=-1&listUrl=%2FMobile%2FIntroduce%2FHealth%2FSearch

%2Findex%2Easp%3FCateid%3D290%26p%3D136
3. 서울의료원, 심장질환
https://seoulmc.or.kr/pms/contents/contents.do?contseqn=371&decorator=pmsweb&menucdv=01020404
4. 로지스틱 회귀분석을 이용한 중소기업 기술 보호 요인 분석, 한국전자거래학회지, 2015년, pp.1 –10,
5. Regression methods for analyzing the risk factors for a life style disease among the young population of India, B. Ismail and Manjula Anil
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4310968/
6. Relation of Helicobacter pylori infection and coronary heart disease. M. A. Mendall, P. M. Goggin, N. Molineaux, J. Levy, T. Toosy, D. Strachan, A. J. Camm, T. C. Northfield
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC483719/
7. Comparing performances of logistic regression, decision trees, and neural networks for classifying heart disease patients, IEEE, A. Khemphilla
https://ieeexplore.ieee.org/abstract/document/5643666
8. Hosmer-Lemeshow 검정(1)
https://support.minitab.com/ko-kr/minitab/18/help-and-how-to/modeling-statistics/regression/how-to/fit-binary-logistic-model/interpret-the-results/all-statistics-and-graphs/goodness-of-fit-tests/
9. Hosmer-Lemeshow 검정(2)
https://en.wikipedia.org/wiki/Hosmer%E2%80%93Lemeshow_test
10. ROC Curve
https://losskatsu.github.io/machine-learning/stat-roc-curve/#
11. Cross Validation
https://www.listendata.com/2015/05/two-ways-to-score-validation-data-in.html
12. Stepwise
https://thebook.io/006723/ch08/05/01/
13. 데이터 출처, UCI Machine Learning Repository
https://archive.ics.uci.edu/ml/datasets/heart+disease