회귀분석 및 실습 II

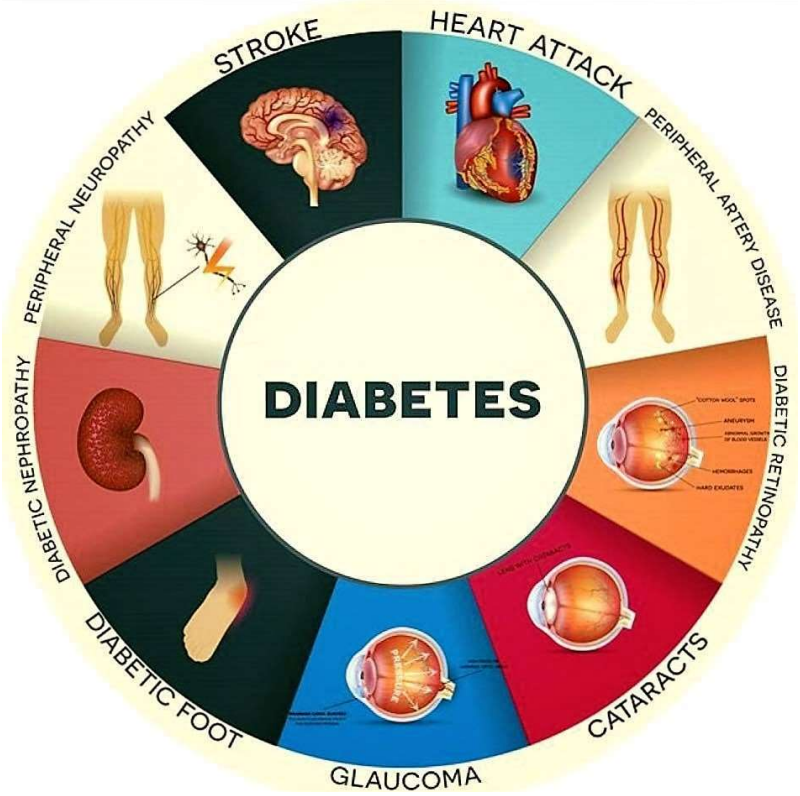
로지스틱 회귀분석을 통한 당뇨병 데이터 분석

2015580023 통계학과 이동균
2017580035 통계학과 이지윤

CONTENT

01 서론

연구주제



당뇨병 발병에 관련된 위험인자 분석

01 서론
데이터설명

- PIMA 인디언 여성에 대한 정보가 담긴 당뇨병 데이터
- 20세 이상의 여성 768명에 대한 데이터
- National Institute of Diabetes and Digestive and Kidney Diseases

01 서론
변수설명

OBS	Pregnancies	Glucose	BloodPressure	SkinThickness	BMI	DiabetesPedigreeFunction	Age	Outcome
1	6	148	72	35	33.6	0.627	50	1
2	1	85	66	29	26.6	0.351	31	0
3	8	183	64	0	23.3	0.672	32	1

변수 종류

Pregnancies – 임신 횟수

Glucose - 당뇨검사 중 2시간 동안의 혈당농도

BloodPressure - 심이지장 혈압

SkinThickness - 삼두근 피부 두께

BMI - BMI 지수

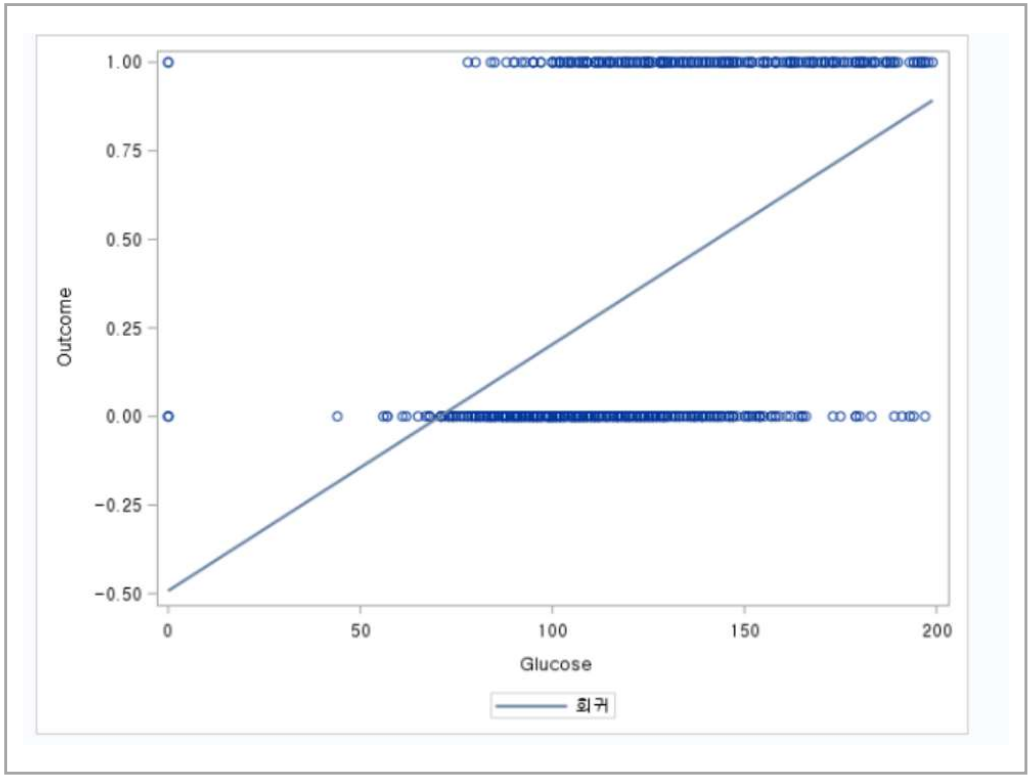
DiabetesPedigreeFunction- 당뇨 혈통함수

Age - 나이

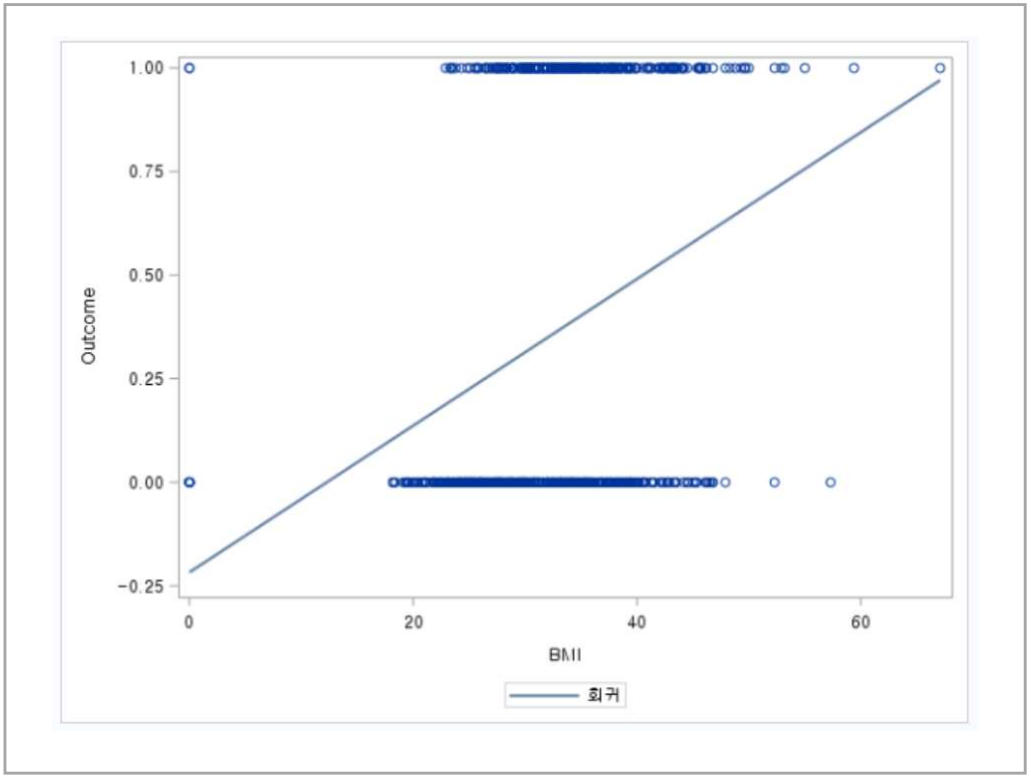
Outcome- 당뇨병 발병 여부 (0- 정상 1-당뇨병)

반응변수: Outcome (당뇨병 발병여부) **종속변수: 나머지 7개의 변수**

X=Glucose Y=Outcome



X=BMI Y=Outcome



선형회귀 분석은 부적절함

로지스틱 회귀분석

(반응변수) Outcome = 0 (정상) or 1 (당뇨병) : *이항변수*

로지스틱 회귀분석이 적절하다.

로지스틱 회귀모형

Logit = $log\left(p / (1-p) \right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n$

상관계수 확인

피어슨 상관 계수, N = 768
H0: Rho=0 가정하에서 Prob > | r|

	Pregnancies	Glucose	BloodPressure	SkinThickness	BMI	DiabetesPedigreeFunction	Age	Outcome
Pregnancies	1.00000	0.12946 0.0003	0.14128 <.0001	-0.08167 0.0236	0.01768 0.6246	-0.03352 0.3535	0.54434 <.0001	0.22190 <.0001
Glucose	0.12946 0.0003	1.00000	0.15259 <.0001	0.05733 0.1124	0.22107 <.0001	0.13734 0.0001	0.26351 <.0001	0.46658 <.0001
BloodPressure	0.14128 <.0001	0.15259 <.0001	1.00000	0.20737 <.0001	0.28181 <.0001	0.04126 0.2534	0.23953 <.0001	0.06507 0.0715
SkinThickness	-0.08167 0.0236	0.05733 0.1124	0.20737 <.0001	1.00000	0.39257 <.0001	0.18393 <.0001	-0.11397 0.0016	0.07475 0.0383
BMI	0.01768 0.6246	0.22107 <.0001	0.28181 <.0001	0.39257 <.0001	1.00000	0.14065 <.0001	0.03624 0.3158	0.29269 <.0001
DiabetesPedigreeFunction	-0.03352 0.3535	0.13734 0.0001	0.04126 0.2534	0.18393 <.0001	0.14065 <.0001	1.00000	0.03356 0.3530	0.17384 <.0001
Age	0.54434 <.0001	0.26351 <.0001	0.23953 <.0001	-0.11397 0.0016	0.03624 0.3158	0.03356 0.3530	1.00000	0.23836 <.0001
Outcome	0.22190 <.0001	0.46658 <.0001	0.06507 0.0715	0.07475 0.0383	0.29269 <.0001	0.17384 <.0001	0.23836 <.0001	1.00000

Glucose, BMI ,Age 순으로 Outcome과의 상관계수가 높게 나타나
이 3개의 변수가 당뇨병에 영향을 줄 것으로 예상됨

① 7개 설명변수 전부를 가지고 로지스틱 회귀분석 실행

- 모형 적합도 검정

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	268.2968	7	<.0001
Score	231.8747	7	<.0001
Wald	168.0359	7	<.0001

귀무가설 (H_0) :β= 0 를 기각한다.

Deviance and Pearson Goodness-of-Fit Statistics				
Criterion	Value	DF	Value/DF	Pr > ChiSq
Deviance	725.1871	760	0.9542	0.8132
Pearson	844.0599	760	1.1106	0.0179

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
4.2583	8	0.8331

귀무가설 (H_0) : 모형이 적합하다 를 기각하지 못한다.

모형은 적합하다고 할 수 있다.

- 계수에 대한 검정

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-8.2689	0.7055	137.3735	<.0001
Pregnancies	1	0.1245	0.0319	15.1943	<.0001
Glucose	1	0.0335	0.00344	94.8129	<.0001
BloodPressure	1	-0.0130	0.00520	6.2572	0.0124
SkinThickness	1	-0.00325	0.00620	0.2749	0.6001
BMI	1	0.0901	0.0150	35.8902	<.0001
DiabetesPedigreeFunc	1	0.9169	0.2980	9.4665	0.0021
Age	1	0.0157	0.00931	2.8342	0.0923

귀무가설 (H_0) : $\beta_i = 0$

SkinThikness 변수는 유의하지 않다.

- 변수선택법 : stepwise selection

Note: No (additional) effects met the 0.1 significance level for entry into the model.

Summary of Stepwise Selection

Step	Effect Entered	Effect Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
1	Glucose		1	1	167.1922		<.0001
2	BMI		1	2	34.3033		<.0001
3	Pregnancies		1	3	27.3305		<.0001
4	DiabetesPedigreeFunction		1	4	9.6773		0.0019
5	BloodPressure		1	5	5.8123		0.0159
6	Age		1	6	3.1493		0.0760

SkinThickness 변수 포함 되지 않음

② SkinThickness 제외한 설명변수 가지고 로지스틱 회귀분석

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	268.0222	6	<.0001
Score	231.7730	6	<.0001
Wald	167.8715	6	<.0001

Deviance and Pearson Goodness-of-Fit Statistics				
Criterion	Value	DF	Value/DF	Pr > ChiSq
Deviance	725.4617	761	0.9533	0.8182
Pearson	838.1838	761	1.1014	0.0267

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
5.2716	8	0.7282

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-8.2398	0.7020	137.7831	<.0001
Pregnancies	1	0.1249	0.0320	15.2658	<.0001
Glucose	1	0.0335	0.00344	94.7943	<.0001
BloodPressure	1	-0.0135	0.00511	6.9537	0.0084
BMI	1	0.0877	0.0143	37.7617	<.0001
DiabetesPedigreeFunc	1	0.8961	0.2949	9.2368	0.0024
Age	1	0.0163	0.00924	3.1232	0.0772

적합도 검정 결과: 모형이 적합하다.

Wald 검정 통계량 : 6개의 설명변수가 유의하다.

-회귀계수 추정

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-8.2398	0.7020	137.7831	<.0001
Pregnancies	1	0.1249	0.0320	15.2658	<.0001
Glucose	1	0.0335	0.00344	94.7943	<.0001
BloodPressure	1	-0.0135	0.00511	6.9537	0.0084
BMI	1	0.0877	0.0143	37.7617	<.0001
DiabetesPedigreeFunc	1	0.8961	0.2949	9.2368	0.0024
Age	1	0.0163	0.00924	3.1232	0.0772

$$logit = log(\frac{\pi(x)}{1-\pi(x)}) = -8.2398 + 0.1249 pregnancies + 0.0335 Glucose - 0.0135 BloodPressure$$
$$+ 0.0877 BMI + 0.8961 DiabetesPedigreeFunction + 0.0163 Age$$

- 오즈

$$\text{Odds} = \frac{\pi(x)}{1-\pi(x)}$$

$$\frac{\frac{\pi(x_i+1)}{1-\pi(x_i+1)}}{\frac{\pi(x_i)}{1-\pi(x_i)}} = e^{\beta_i}$$

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Pregnancies	1.133	1.064	1.206
Glucose	1.034	1.027	1.041
BloodPressure	0.987	0.977	0.997
BMI	1.092	1.062	1.123
DiabetesPedigreeFunction	2.450	1.375	4.367
Age	1.016	0.998	1.035

나머지 설명변수의 값들이 고정되었다는 가정하에
설명변수가 한 단위 증가 시 반응변수의 odds 증가량을 의미한다.

수치적으로 해석한 결과를 보면

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Pregnancies	1.133	1.064	1.206
Glucose	1.034	1.027	1.041
BloodPressure	0.987	0.977	0.997
BMI	1.092	1.062	1.123
DiabetesPedigreeFunction	2.450	1.375	4.367
Age	1.016	0.998	1.035

값이 높을수록 유병률이 높아지는 변수
: 임신횟수, 혈당농도, BMI수치,
당뇨혈통함수, 나이

값이 낮을수록 유병률이 높아지는 변수
:혈압수치

변수	N	평균	표준편차	최솟값	최댓값
Pregnancies	768	3.8450521	3.3695781	0	17.0000000
Glucose	768	120.8945313	31.9726182	0	199.0000000
BloodPressure	768	69.1054688	19.3558072	0	122.0000000
SkinThickness	768	20.5364583	15.9522176	0	99.0000000
BMI	768	31.9925781	7.8841603	0	67.1000000
DiabetesPedigreeFunction	768	0.4718763	0.3313286	0.0780000	2.4200000
Age	768	33.2408854	11.7602315	21.0000000	81.0000000

하지만, 설명변수 데이터들의 scale의 차이 때문에
어떤 변수가 가장 큰 영향을 미친다고 결론 내기 어려움

→ 설명변수 표준화

② 표준화 시킨 후 로지스틱 회귀분석

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.8720	0.0969	81.0640	<.0001
Pregnancies	1	0.4209	0.1077	15.2658	<.0001
Glucose	1	1.0708	0.1100	94.7943	<.0001
BloodPressure	1	-0.2610	0.0990	6.9537	0.0084
BMI	1	0.6912	0.1125	37.7617	<.0001
DiabetesPedigreeFunc	1	0.2969	0.0977	9.2368	0.0024
Age	1	0.1920	0.1086	3.1232	0.0772

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Pregnancies	1.523	1.233	1.882
Glucose	2.918	2.352	3.620
BloodPressure	0.770	0.634	0.935
BMI	1.996	1.601	2.489
DiabetesPedigreeFunction	1.346	1.111	1.630
Age	1.212	0.979	1.499

$$logit\left(\frac{\pi(x)}{1-\pi(x)}\right) = -0.8720 + 0.4209 pregnancies + 1.0708 Glucose - 0.2610 BloodPressure$$
$$+ 0.6912 BMI + 0.2969 DiabetesPedigreeFunction + 0.1920 Age$$

Glucose , BMI, Pregnancies 순으로 영향을 많이 끼친다.

분석 결과

- 삼두근 피부 두께는 당뇨병 발병여부에 영향을 끼치지 않음.
- 임신횟수, 혈당농도, 혈압, BMI, 당뇨혈통함수,나이는 영향을 끼침.
- 가장 당뇨병 발병여부에 가장 영향을 많이 끼치는 변수 3개는 혈당농도, BMI수치, 임신횟수 라고 결론을 내릴 수 있다.

03 결론
분석의 한계점 및 제안사항

분석의 한계점

1. 분석에 사용한 데이터는 건강 수치에 관한 설명변수만 포함
2. Pima Indian 을 대상으로 한 데이터이므로
   아시아인 우리나라 사람들에게도 같은 결과를 가질 것이라 단정하기 어려움

제안사항

아시아인 또는 우리나라 국민들을 대상으로 운동빈도, 스트레스 정도, 가구 소득 등 이러한 분야의 요인들의 정보를 획득하여 분석하면 개선될 것이라 생각한다.

감사합니다.