선형회귀분석/로지스틱 회귀분석/ROC 분석

신 민 호
전남대학교 의과대학 예방의학교실

변수란?

- 변수(Variable)
  - 개체에 따라 변화가 가능한 상태 또는 특성
  - 예)성별, 연령, 키
- 변수의 종류
  - 독립변수(independent variable)
  - 종속변수(dependent variable)
- 척도의 종류
  - 명목척도
  - 순위척도
  - 간격척도
  - 비율척도

통계분석방법

종속변수	독립변수	통계분석방법
범주형	범주형	Chi-square test
	연속형 or 범주형	Logistic regression
연속형	범주형(2개 이상의 범주)	t-test, pared t-test
	범주형(3개 이상의 범주)	ANOVA
	연속형 or 범주형	Linear regression
연속형(생존기간)	연속형 or 범주형	Survival analysis

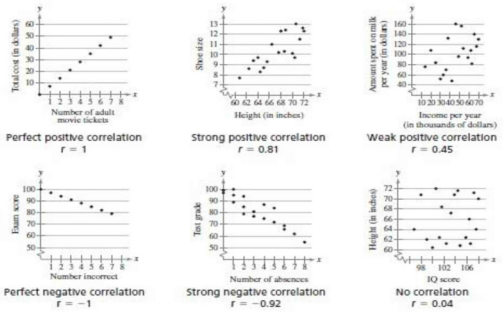
Correlation

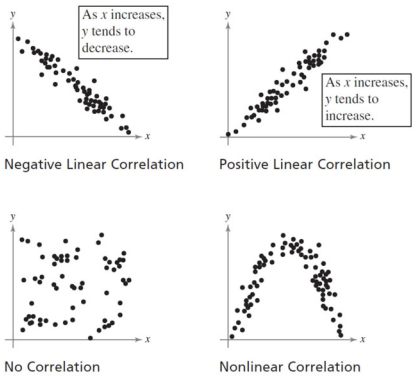
- 상관분석(correlation analysis)
  - 두 변수 x(독립변수)와 y(종속변수)간의 관련성 정도를 측정하기 위한 분석방법
  - x와 y는 수치형 변수(예. 키, 몸무게)
  - 산점도를 이용하여 이차원 그림으로 표시 가능
- 상관계수(correlation coefficient, r)
  - 변수 간의 선형관계와 강도를 나타내는 척도

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n\sum x^2 - (\sum x)^2}\sqrt{n\sum y^2 - (\sum y)^2}}$$

  - r 값의 범위는 -1에서 +1사이
  - r값의 부호는 양의 관계인지, 음의 관계인지 알려줌
  - r값의 크기는 점들이 직선에 얼마나 가까운지 알려줌
  - r값이 -1이나 +1에 가까울수록 선형적인 관련성 정도 증가
  - X와 y의 상관이 반드리 x와 y간의 '인과관계'를 의미하는 것이 아님
  - r^2은 y의 변동량 중 x와 y간의 선형관계로 설명될 수 있는 비율

Correlation





Linear regression

- 회귀분석
  - 독립변수에 대한 종속변수의 결과
- 결정계수(R^2)
  - 추정된 회귀직선에 의해 변동이 설명되는 정도

Total deviation $= y_i - \bar{y}$

Explained deviation $= \hat{y}_i - \bar{y}$

Unexplained deviation $= y_i - \hat{y}_i$

$$r^2 = \frac{\text{Explained variation}}{\text{Total variation}}$$

예)상관계수 r의 값이 0.90일 때, 결정계수는?

r^2=0.902
=0.81

Simple linear regression

- 한 개의 독립변수에 대한 종속변수의 결과 예측
- y=mx + b (m:기울기, b:절편)

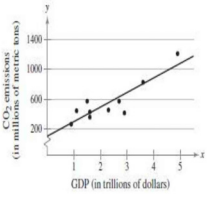
$$m = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2} \qquad b = \bar{y} - m\bar{x} = \frac{\sum y}{n} - m\frac{\sum x}{n}$$

GDP (trillions of $), x	CO_2 emissions (millions of metric tons), y
1.6	428.2
3.6	828.8
4.9	1214.2
1.1	444.6
0.9	264.0
2.9	415.3
2.7	571.8
2.3	454.9
1.6	358.7
1.5	573.5

$$m = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2} = \frac{10(15{,}573.71) - (23.1)(5554)}{10(67.35) - 23.1^2} = \frac{27{,}439.7}{139.89} \approx 196.151977$$

$$b = \bar{y} - m\bar{x} \approx \frac{5554}{10} - (196.151977)\frac{23.1}{10} = 555.4 - (196.151977)(2.31) \approx 102.2889$$

$$\hat{y} = 196.152x + 102.289.$$



Multiple linear regression

- 두 개 이상의 독립변수에 대한 종속변수의 결과 예측

$$\hat{y} = b + m_1x_1 + m_2x_2 + m_3x_3 + \cdots + m_kx_k$$

- 예)연구자들은 A회사의 직원 급여는 고용기간, 이전의 경험, 교육이 어떤 관련을 갖는지 알고자 한다. 연구자들은 8명의 고용자들을 선택했고, 그 결과는 아래의 표와 같다.

Employee	Salary, y	Employment (in years), x_1	Experience (in years), x_2	Education (in years), x_3
A	57,310	10	2	16
B	57,380	5	6	16
C	54,135	3	1	12
D	56,985	6	5	14
E	58,715	8	8	16
F	60,620	20	0	12
G	59,200	8	4	18
H	60,320	14	6	17

Predictor	Coef		SE Coef	T	P
Constant	49764	b	1981	25.12	0.000
x1	364.41	m_1	48.32	7.54	0.002
x2	227.6	m_2	123.8	1.84	0.140
x3	266.9	m_3	147.4	1.81	0.144

S = 659.490 RSq = 94.4% RSq(adj) = 90.2%

$$\hat{y} = 49{,}764 + 364x_1 + 228x_2 + 267x_3$$

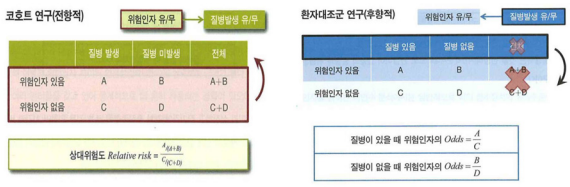
Odds ratio

상대위험도(Relative Risk)

- 위험인자가 없는 경우에 비해 위험인자가 있을 때 질병이 발생할 상대적 위험도
- 코호트 연구에서 사용

교차비(Odds Ratio)

- 질병이 있는 경우 위험인자 유무의 비와 질병이 없는 경우 위험인자 유무의 비의 비
- 환자-대조군 연구에서 주로 사용



Odds ratio

- 예) 흡연자 100명과 비흡연자 100명을 20년 관찰하였더니, 폐암발생이 흡연자에서는 30명, 비흡연자에서는 5명이 발생하였다. 이때의 RR과 OR은?

	폐암(+)	폐암(-)	합계
흡연	30	70	100
비흡연	5	95	100

- 상대위험도(RR)
  - 흡연자 중에서 폐암이 발생할 확률=30/100=0.3
  - 비흡연자 중에서 폐암이 발생할 확률=5/100=0.05
  - RR=0.3/0.05=6
- 교차비(OR)
  - 폐암환자 중에 흡연의 odds = 30/5
  - 폐암 환자가 아닌 사람 중에 흡연의 odds=70/95
  - OR=(30/5)/(70/95)=(30*95)/(5*70)=8.1

Odds ratio

- 상대위험도와 교차비의 관계

	질병 발생	질병 미발생	전체
위험인자 있음	p_2	$1-p_2$	1
위험인자 없음	p_1	$1-p_1$	1

교차비 (Odds ratio) $\frac{\frac{p_2}{1-p_2}}{\frac{p_1}{1-p_1}} = \frac{p_2(1-p_1)}{p_1(1-p_2)} \cong \frac{p_2}{p_1} =$ 상대위험도 (Relative risk)

만약 질환 발생이 드물어 p_1, p_2가 매우 작다면
$\frac{1-p_1}{1-p_2} \cong 1$ 에 근사한다고 가정할 수 있다.

Logistic regression

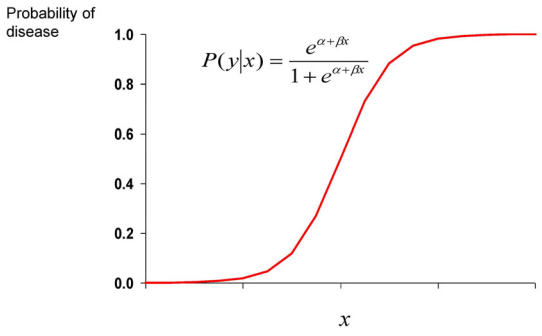
- 로지스틱 회귀분석(logistic regression analysis)
  - 특정 질병의 유무(종속변수:이분형)에 영향을 미치는 요인(독립변수)을 밝히고자 할 때 사용
  - 여러 위험인자들이 관련되는 정도를 하나의 모형을 설명하기에 적합

LOGISTIC FUNCTION: $f(z)=1/[1+\exp(-z)]$

LOGISTIC MODEL: $P(\mathbf{X})=1/[1+\exp[-(\alpha+\Sigma\beta_i X_i)]]$

LOGIT TRANSFORMATION: $\text{logit } P(\mathbf{X})= \alpha+\Sigma\beta_i X_i$

Logistic function



Logistic transformation

$$P(y|x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$$

$$\ln\left[\frac{P(y|x)}{1-P(y|x)}\right] = \alpha + \beta x$$

logit of $P(y|x)$

Interpretation of coefficient b

	Exposure x	
Disease y	yes	no
yes	$P(y\|x=1)$	$P(y\|x=0)$
no	$1-P(y\|x=1)$	$1-P(y\|x=0)$

$$\frac{P}{1-P} = e^{\alpha+\beta x}$$

$$Odds_{d|e} = e^{\alpha+\beta} \qquad OR = \frac{e^{\alpha+\beta}}{e^{\alpha}} = e^{\beta}$$

$$Odds_{d|\bar{e}} = e^{\alpha} \qquad \ln(OR) = \beta$$

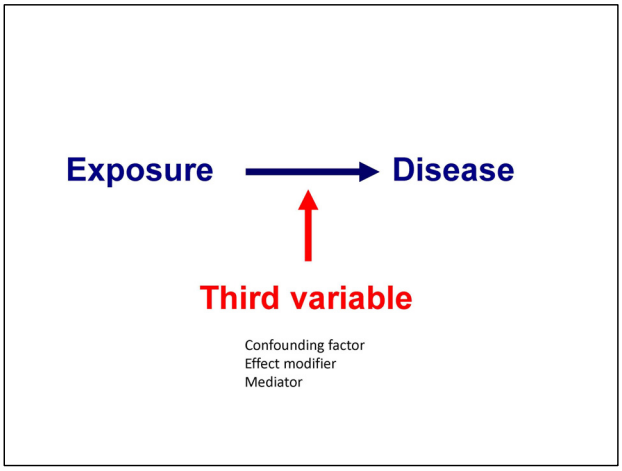
Interpretation of coefficient b

$$\ln\left(\frac{P}{1-P}\right) = \alpha + \beta_1 \times \text{Age} = -0.841 + 2.094 \times \text{Age}$$

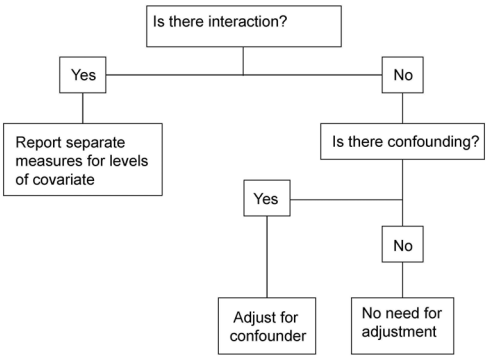
	Coefficient	SE	Coeff/SE
Age	2.094	0.529	3.96
Constant	-0.841	0.255	-3.30

OR $= e^{2.094} = 8.1$

95% CI $= e^{(2.094 \pm 1.96 \times 0.529)} = 2.9, 22.9$



Interaction and Confounding



Confounding

	Malaria Cases	Controls
Males	88	68
Females	62	82
	150	150

OR=1.71

Outdoor occupation

	Cases	Controls
Males	53	15
Females	10	3
	63	18

OR=1.06

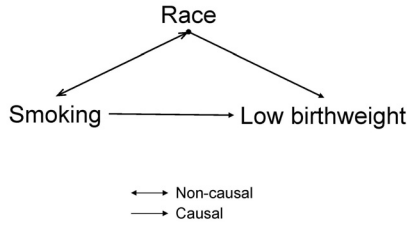
Indoor occupation

	Cases	Controls
Males	35	53
Females	52	79
	87	132

OR=1.00

Adjusted OR=1.01

Confounding

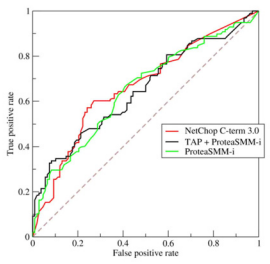


Ways to control for confounding

- During the design phase of the study:
  - Randomized trial
  - Matching
  - Restriction
- During the analysis phase of the study:
  - Stratification
  - Adjustment
- Newer approaches
  - Graphical approaches using DAGs
  - Propensity scores
  - Instrumental variables
  - Marginal structural models

Receiver operating characteristic (ROC)

- plot of the sensitivity vs. (1 - specificity) for a binary classifier system as its discrimination threshold is varied.



ROC 분석의 활용

- 진단검사의 타당성 평가
- 2가지 이상 진단검사 방법들의 유용성 평가
- 진단검사의 구분점(cut-off point) 결정
- 로지스틱 모델의 적합도 평가(C statistic)

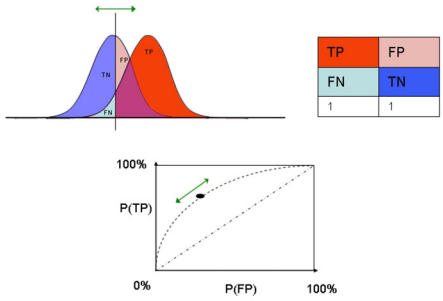
The measures of diagnostic accuracy

Predicted Labels	Actual Labels		Total
	Positive ($Y = 1$)	Negative ($Y = 0$)	
Positive ($\hat{Y} = 1$)	TP	FP	TP + FP
Negative ($\hat{Y} = 0$)	FN	TN	FN + TN
Total	TP + FN	FP + TN	n

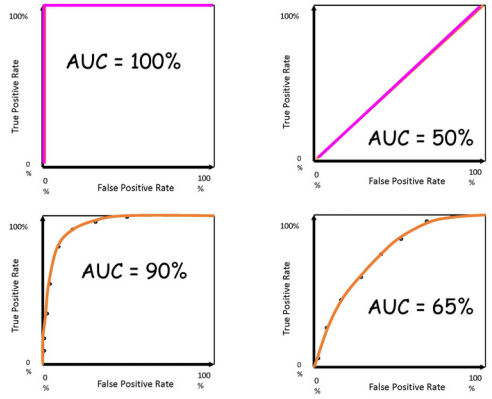
TP: True positive
FP: False positive
TN: True negative
FN: False negative
NPV: Negative predictive value
PPV: Positive predictive value
PLR: Positive likelihood ratio
NLR: Negative likelihood ratio

Sensitivity $= TP/(TP + FN)$
Specificity $= TN/(FP + TN)$
$PPV = TP/(TP + FP)$
$NPV = TN/(TN + FN)$
$PLR = Sensitivity/(1 - Specificity)$
$NLR = (1 - Sensitivity)/Specificity$

The R Journal, 8(2):213-230, 2016



AUC for ROC curves



Grading guidelines for AUC values

- 0.90 - 1.00 = excellent discrimination
- 0.80 - 0.90 = good discrimination
- 0.70 - 0.80 = fair discrimination
- 0.60 - 0.70 = poor discrimination
- 0.50 - 0.60 = failed discrimination

Determination of optimal cut-off value

- Youden's index*(J)*
  = maximum(sensitivity + specificity-1)
- the point on the ROC curve closest to (0,1)
  - Minimum value of the square root of [(1- sensitivity)^2-(1-specificity)^2]

Methods for estimating ROC curve and AUC

- Bamber (1975)
- Hanley and McNeil(1982)
- Metz et al(1984)
- DeLong, DeLong, and Clarke-Pearson (1988)

ROC softwares

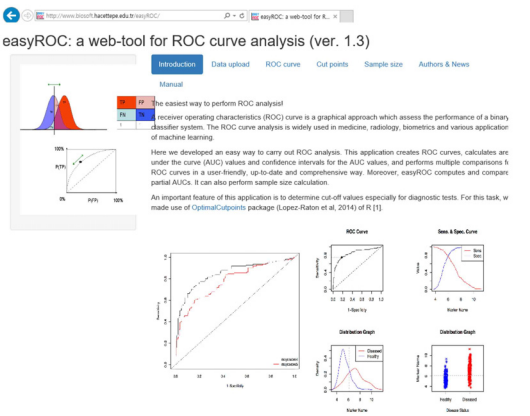
- 범용프로그램
  - SPSS, SAS, STATA
- 전용프로그램
  - AccuRoc, Analyse-It, CMDT, GraphROC, MedCalc, mROC, ROCKIT
- R packages
  - pROC, ROCR, ROC
- Web based program
  - easyROC

ROC softwares comparison

	IBM SPSS	Stata	MedCalc	ROC	ROCR	pROC	easyROC
Plots	Yes	Yes	Yes*	Yes	Yes*	Yes*	Yes*
Conf. intervals	Yes	Yes*	Yes	Yes	Yes	Yes*	Yes*
pAUC	No	Yes	No	Yes	Yes	Yes*	Yes*
Statistical tests	No	Yes	No	Yes	Yes	Yes*	Yes*
Diagnostic measures	No	Yes	Yes	No	Yes*	Yes	Yes
Multiple comp.	No	Yes	Yes*	No	No	Yes*	Yes
Cutpoints	No	Yes	Yes	No	No	Yes	Yes*
Sample size	No	No	Yes	No	No	Yes	Yes*
Free license	No	No	No	Yes*	Yes*	Yes*	Yes*
Open source	No	No	No	Yes*	Yes*	Yes*	Yes*
Web-tool access	No	No	No	No	No	No	Yes*
User interface	Yes	Yes*	Yes*	No	No	Yes*	Yes*

* Comprehensive ones.

The R Journal, 8(2):213-230, 2016



실습예제 1

NAME: LOW BIRTH WEIGHT DATA (LOWBWT.DAT
SOURCE: Hosmer, D.W., Lemeshow, S. and Sturdivant, R.X. (2013) Applied Logistic Regression: Third Edition.

```
LIST OF VARIABLES:

Columns     Variable                                              Abbreviation
-----------------------------------------------------------------------------
2-4         Identification Code                                   ID

10          Low Birth Weight (0 = Birth Weight >= 2500g,          LOW
                              1 = Birth Weight < 2500g)

17-18       Age of the Mother in Years                            AGE

23-25       Weight in Pounds at the Last Menstrual Period         LWT

32          Race (1 = White, 2 = Black, 3 = Other)                RACE

40          Smoking Status During Pregnancy (1 = Yes, 0 = No)     SMOKE

48          History of Premature Labor (0 = None  1 = One, etc.)  PTL

55          History of Hypertension (1 = Yes, 0 = No)             HT

61          Presence of Uterine Irritability (1 = Yes, 0 = No)    UI

67          Number of Physician Visits During the First Trimester FTV
                 (0 = None, 1 = One, 2 = Two, etc.)

73-76       Birth Weight in Grams                                 BWT
-----------------------------------------------------------------------------
```

실습예제 2

Supplementary 2: Non-alcoholic fatty liver disease (NAFLD) data (Çelikbilek et al., 2014).

Grup	mir197	mir146b	mir181d	mir99a	Grup	mir197	mir146b	mir181d	mir99a
1	0.921	0.687	0.474	−0.941	0	1.214	1.122	0.882	1.610
1	0.967	1.059	0.474	0.575	0	1.401	0.148	0.444	0.625
1	0.854	1.105	0.722	0.936	0	0.494	−0.179	1.386	0.134
1	−1.088	−1.353	−0.577	−1.077	0	1.608	1.386	2.242	0.926
1	0.107	0.515	−0.286	0.560	0	1.274	1.609	0.769	1.108
1	0.547	1.191	0.583	1.119	0	0.827	1.128	0.452	0.374
1	−1.081	−1.445	−1.303	−1.202	0	−0.147	−0.545	0.878	0.044
1	−1.081	−1.308	−1.276	−1.066	0	0.353	0.320	−0.225	0.367
1	0.841	0.463	−0.290	0.747	0	−1.635	−0.677	−0.838	−0.543
1	−1.188	−0.975	−1.407	−2.123	0	1.848	1.523	1.712	0.940
1	−1.014	−0.649	−1.194	−1.786	0	0.987	0.606	0.626	0.542
1	−1.081	−1.256	−1.229	−0.679	0	0.020	0.503	0.600	0.367
1	−1.295	−1.204	−1.607	−2.216	0	1.061	1.518	1.217	0.209
1	−1.081	−1.268	−0.829	−0.658	0	0.474	0.572	0.292	0.786
1	−1.081	−1.365	−1.376	−1.457	0	−0.868	−0.505	−0.408	−0.117
1	−1.081	−1.371	−0.812	−1.804	0	−0.414	−0.259	0.665	0.363
1	−1.081	−0.769	−1.359	−0.156	0	0.394	0.417	1.000	0.130
1	0.854	1.243	0.444	1.460	0	0.941	0.543	0.431	1.083
1	−1.074	−1.365	−1.572	−0.339	0	−0.387	−0.202	−0.568	0.345
1	−0.634	−0.276	−0.130	−0.081	0	−0.674	−0.689	0.995	0.893

The R Journal, 8(2):213-230, 2016