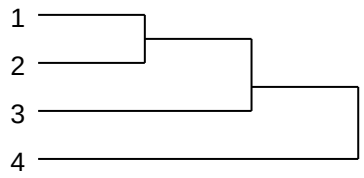
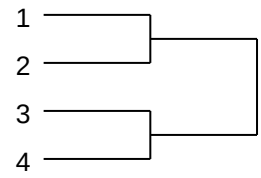
2 장에서는 두 범주형 변수의 관계를 (독립성) 분석하는 IxJ 분할표 분석을 살펴보았다. 이를 정리하면

- 대표본 (기대 빈도가 5 미만인 셀이 없거나 전체 셀 개수의 20% 넘지 않으면)인 경우 두 변수의 독립성 검정은 Pearson χ^2검정, LR Chi-square(G^2) 방법을 사용한다.
- 소표본 분할표는 Fisher's exact test 를 한다.
- 2x2 분할표는 두 모집단 비율 차이 검정, Odds Ratio θ 검정, Chi-square 검정을 할 수 있다.
- Chi-square 검정은 수집 자료가 어떤 분포를 따르는지 적합성에 (Goodness of fit) 이용할 수 있다.
- 두 범주형 변수가 순서형이면 선형 상관 분석을 할 수 있다. Pearson cross moment (검정은 Mantel-Haenszel 검정), 이와 유사한 Phi-coefficient, Gamma (γ), Kendall τ 등이 있다.
- 설명변수가 순서형이고 반응변수가 이진형인 (binary) 경우 Cochran-Armitage Trend 방법을 사용하여 반응변수의 성공률의 직선 변화를 살펴볼 수 있다.
- Chi-square 분할표를 sub 분할표로 분할하여 관심 있는 범주들간의 상관 관계를 살펴볼 수 있다. 분할표 분할은 원래 분할표가 유의한 경우 (Chi-square 통계량이 기각값보다 커 귀무가설을 기각) 사용해야 한다. 다음은 분할 방법을 요약한 것이다.

Hermit Contrast 방법	Generating Hierarchical Structure
	

분할표 검정은 두 범주형 변수의 상관 관계를 (association) 분석하는데 사용할 수 있으나 변수가 3 개 이상이고 설명 변수들의 반응 변수에 대한 효과를 분석하려면 모형화가 (modeling) 필요하다. 이 모형들을 Generalized Linear Model (GLM: 일반화 선형 모형)이라 한다. GLM 은 반응변수가 측정형인 경우 분석하는 회귀분석 (Regression), 분산분석 (ANOVA: Analysis of Variance)은 물론 반응변수가 범주형인 경우 사용하는 Logistic, Log-Linear Model 까지 포함하고 있다.

3.1. Generalized Linear Model

Neder & Wedderburn(1972) 의해 제안된 모형을 일반화 한 GLM 은 3 가지 성분에 (component) 의해 정의된다. 1)random component: 반응변수의 확률분포함수 인식 2)systematic component: 설명변수의 (predictor 라고도 함) 선형 함수 규정 3) link: systematic 성분과 random 성분의 기대값 사이의 함수 관계를 표현.

3.1.1. GLM 의 성분 (component)

Random component

랜덤 성분은 natural exponential family(자연 대수 군집)의 분포로부터 추출된 서로 독립인 관측치 $Y=(Y_1,Y_2,...Y_n)$ 로 구성되어 있다. 즉 각 관측치 Y_i 의 확률밀도함수는 다음과 같다.

$$f(y_i;\theta_i)=a(\theta_i)b(y_i)\exp[y_iQ(\theta_i)],\ \theta_i \text{ 는 모수(parameter)}$$

Poisson 분포 (분할표의 셀의 관측빈도), Binomial 분포 (성공 회수), Standard Normal 분포 (일반 회귀분석)가 natural exponential family 에 속한다. 모수(θ_i)의 값은 관측치마다 변할 수 있다. $Q(\theta_i)$ 를 자연 대수 모수라 한다.

Systematic component

설명변수로 구성된 행렬 X (자료 행렬 data matrix 혹은 design matrix), 모형의 모수 벡터를 $\underline{\beta}$ 라 하자. 다음의 linear predictor (선형 예측치)가 GLM 의 systematic 성분이다.

$$\underline{\eta}=X\underline{\beta}\Rightarrow\eta_i=\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+...+\beta_px_{ip}=\sum_j\beta_jx_{ij}\quad \text{for}\ \ i=1,2,....n$$

Link component

Random 성분과 systematic 성분을 연결하는 성분을 의미한다. 관측치 Y_i 의 기대치를 $\mu_i=E(Y_i)$ 라 하자. 이 때 μ_i 는 $\eta_i=g(\mu_i)$ 식에 의해 η_i 와 연결된다. link 함수 g 는 미분 가능한 단조 함수이다. (monotonic differential function)

$$g(\mu_i)=\sum_j\beta_jx_{ij}$$

연결 함수의 간단한 형태는 identity link (항등 연결)인 $g(\mu)=\mu$ 이다. 이것은 평균 반응 모형이며 일반적인 회귀 모형이다.

$$E(y_i)=\mu_i=\sum_j\beta_jx_{ij}=\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+...+\beta_px_{ip}$$

평균을 자연 대수 모수로 변환하는 연결 함수를 Canonical Link 라 한다. 즉 Canonical Link 에서는 $g(\mu_i)=Q(\theta_i)=\sum_j\beta_jx_{ij}=\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+...+\beta_px_{ip}$ 이다. Canonical Link 가 가장 일반적인 연결 함수이다.

3.1.2. Logit model

범주가 2 개인 경우 (이진, binary 예; 성공/실패, 범주 값을 0, 1 로 표시할 수 있다. 이진 변수는 성공 확률이 $\Pr(Y=1)=\pi$ 인 Bernoulli 분포를 따른다) 범주형 확률 밀도 함수는

$$f(y_i;\pi_i)=\pi_i^{y_i}(1-\pi_i)^{1-y_i}=(1-\pi_i)[\pi_i/(1-\pi_i)]^{y_i}=(1-\pi_i)\exp[y_i\ln(\frac{\pi_i}{1-\pi_i})]\text{ : NE Family}$$

자연 대수 모수 $Q(\theta_i)=\ln(\frac{\pi_i}{1-\pi_i})$ 은 odds ratio 의 ln 값으로 π 의 Logit 이라 한다. 이 Logit 연결을 사용하는 GLM 을 Logit 모형이라 한다.

$$\ln(\frac{\pi_i}{1-\pi_i})=\sum_j\beta_j x_{ij}=\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+...+\beta_p x_{ip}$$

3.1.3. Log Linear model

분할표의 셀의 빈도 n_i 는 Poisson 분포를 따른다고 가정한다. 셀 n_i 의 기대 빈도를 $E(n_{ij})=m_i$ 라 하면 n_i 의 확률밀도함수는

$$f(n_i;m_i)=\frac{\exp(-m_i)(m_i)^{n_i}}{n_i!}=\exp(-m_i)(\frac{1}{n_i!})\exp[n_i\ln(m_i)]\text{ : NE Family}$$

자연 대수 모수 $Q(\theta_i)=\ln(m_i)$ 이다.

$$\ln(m_i)=\sum_j\beta_j x_{ij}=\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+...+\beta_p x_{ip}$$

3.1.4. GLM 모형 분석 방법

Random 성분, Systematic 성분, Link 함수에 의해 GLM 분석 방법을 정리하면 다음과 같다.

Random 성분 (반응변수)	연결 함수	Systematic 성분 (설명변수)	Model (분석 방법)
Normal	Identity (항등)	연속 (측정)	Regression
Normal	Identity	범주	ANOVA
Normal	Identity	Mixed (연속+범주)	Regression with Indicator ANCOVA
Binomial	Logit	Mixed	Logistic Regression
Poisson	Log	Mixed	Log-Linear
Multinomial	Generalized Logit	Mixed	Multinomial response

전통적인 분석 방법은 반응변수를 변환하여 일정한 분산을 갖는 정규분포에 근사 시켜 최소 자승 방법을 (Least Square Method) 사용한다(일반적인 회귀분석). 이와는 대조적으로 GLM 에서는 반응변수가 더 이상 정규분포가 아니거나 근사하지 않으므로 추정 방법은

최소 자승법과 다르다. GLM 에서는 연결 함수의 선택과 Random 성분의 선택은 별개이고 log 우도 함수는 (likelihood function) strictly concave 하므로 ML estimate(최대 우도 추정치)가 존재한다. ML 추정치는 Fisher's scoring 이라는 iteration algorithm 에 의해 계산된다. 이 추정치 계산은 연결 함수나 Random 성분의 확률변수 선택에 상관없이 적용될 수 있다.

3.2. Logistic Regression

반응 변수 Y 를 이진 변수라 하자. 예를 들어 성공/실패, 취업/미취업, 만족/불만족 등 범주가 2 개인 변수를 이진 변수라 하며 자료 코딩 시는 0, 1 로 한다. 이진 변수의 확률 밀도 함수는 성공 확률이 $\Pr(Y=1)=\pi$ 인 Bernoulli 시행이다. 그러므로 Y 의 기대치는 π 이고 분산은 $\pi(1-\pi)$ 이다. 독립인 이진 변수에 의한 성공 횟수($\sum Y_i$)의 분포는 Binomial 분포이다.

3.2.1. Linear Probability Model (선형 확률 모형)

이진 반응변수에 대해 선형 모형 $E(Y)=\pi(x)=\beta_0+\beta_1 x$ 을 선형 확률 모형이라 한다. 이를 Identity Link (항등 연결)이라 한다. 성공 확률이 설명변수 x 의 값에 따라 선형적(linear)으로 변한다. 이 모형은 이항 랜덤 성분과 항등 연결 함수를 갖는 GLM 이다. 이 모형은 x 의 큰 값, 작은 값에 따라 성공 확률(π)이 음수이거나 1 이상의 값으로 추정될 수 있는 구조적 문제를 갖고 있다.

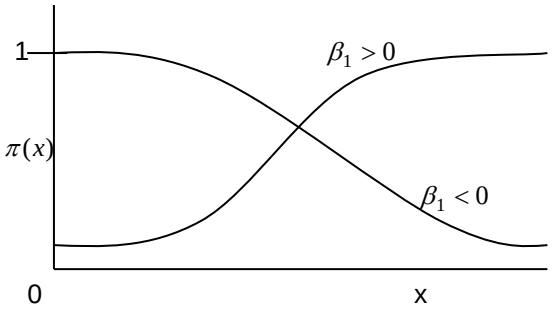
분산 $V(Y)=\pi(1-\pi)$ 은 일정하지 않고 성공 확률이 0 이나 1 로 가까워짐에 따라 분산은 0 에 가까워진다. 그러므로 더 이상 일반 추정치는 MVLUE 는 아니다.

3.2.2. Logistic Regression Model

선형 확률 모형은 구조적 문제가 있다. 성공 확률(π)은 x 와 선형적인 관계가 있다기보다는 비선형 가정할 수 있다. x 의 변화량은 π 가 0 이나 1 에 가까이 있을 때 영향력이 적을 것이다. 이 관계를 S-형태 곡선으로 나타낼 수 있을 것이다.

$$\log it(\pi(x))=\ln\left(\frac{\pi(x)}{1-\pi(x)}\right)=\beta_0+\beta_1 x$$

$$\pi(x)=\frac{\exp(\beta_0+\beta_1 x)}{1+\exp(\beta_0+\beta_1 x)}$$



위의 변환을 log odds 변환 모형을 Logit 모형이라 한다. 즉 x 가 ∞ 로 감에 따라 성공 확률은 β_1 의 부호에 따라 0($\beta_1 < 0$) 혹은 1($\beta_1 > 0$)에 근사한다. 만약 β_1 가 0 이면 반응변수는 설명변수 x 와 독립이다(영향을 받지 않는다). $\frac{\partial \pi(x)}{\partial x} = \beta_1 \pi(x)[1-\pi(x)]$ 이므로 $\pi(x) = 1/2$ 에서 기울기가 가장 급하며 (크며) $|\beta_1|$ 이 클수록 기울기가 급해진다.

Inference (추정)

Logit 모형에서 회귀 계수($\beta_i, i = 1,2,..., p$)의 추정 및 검정 MLE (Maximum Likelihood Estimate)에 대한 Wald (1943) 연구로부터 모수에 대한 대표본 신뢰구간은 다음과 같다.

$\beta_i \pm z_{\alpha/2} ASE(\beta_i)$: ASE = Asymptotic Standard Error (근사 표준 오차)

$\beta_* = (\beta_1, \beta_2, .., \beta_q)'$ 를 모형의 모수 subset 이고 $\beta_* = (\beta_1, \beta_2, .., \beta_q)' = \underline{0}$ 을 검정한다고 하자. (예를 들면 첫번째 설명 변수의 유의성을 검정하려면 $H_0 : \beta_1 = 0$, 만약 2 번째, 3 번째 설명 변수의 유의성을 검정하려면 $H_0 : \beta_2 = \beta_3 = 0$ 이다) L_1 을 Full-모형에서의 로그 우도 함수, L_2 을 Reduced-모형($\beta_* = (\beta_1, \beta_2, .., \beta_q)' = \underline{0}$ 라 하고 모형을 축소)에서의 로그 우도 함수라 하면 다음이 성립한다. 이를 GLM 의 Deviance(벗어남)라 정의한다. Reduced 모형의 우도 함수가 Full 모형의 우도 함수 값의 차이가 적으면 귀무가설에서 유의하지 않다고 설정한 설명변수(회귀 계수)는 반응 변수를 유의적으로 설명하지 못하다는 것이다.

GLM 의 Deviance ➔ $-2\ln(\frac{l_2}{l_1}) = -2[\ln l_2 - \ln l_1] = -2[L_2 - L_1] \sim \chi^2(q)$:

위의 결과는 Theorem -2우도비 $\sim \chi^2$ 으로부터 (유사 결과: 페이지 27 참고)

Wald(1943)는 모수 추정치의 대표본 정규 분포 근사 이론에 근거하여 다음을 증명하였다.

Wald 통계량 ➔ $\hat{\beta}_*' (\hat{Cov}(\hat{\beta}_*)) \hat{\beta}_* \sim \chi^2(q)$:

Logit 모형의 계수 추정에 대한 자세한 내용은 Categorical Data Analysis –Alan Agresti (1990), Wiley publication- page 112-117 참고하기 바란다.

3.2.3. Inverse CDF(역함수) Links

페이지 61 에서 성공 확률($\pi(x)$) 함수의 형태는 ($\beta_1 > 0$)인 경우 누적 분포 함수의 (cumulative probability density function)의 형태와 유사하다. 만약 $\beta_1 < 0$ 인 경우는 x 대신 $-x$ 대치하면 같은 곡선을 얻는다.

이 사실을 ($\pi(x) = F(\beta_0 + \beta_1 x)$: F 는 누적확률밀도 함수) 이용하여 누적밀도함수의 역함수를 연결 함수로 갖는 GLM 을 얻을 수 있다.

$$F^{-1}(\pi(x)) = \beta_0 + \beta_1 x$$

$\beta_1 > 0$ 인 경우 logistic 회귀 모형 $\pi(x) = \exp(\beta_0 + \beta_1 x)/[1+\exp(\beta_0 + \beta_1 x)]$ 은 Logistic 분포의 확률밀도함수와 유사하다.

- Logistic 확률밀도함수(pdf) $f(x \mid \mu, \beta) = \frac{1}{\beta} \frac{\exp(-(x-\mu)/\beta)}{[1+\exp(-(x-\mu)/\beta)]^2}$ 평균= μ , 분산= $\frac{\pi^2 \beta^2}{3}$
- Logistic 누적확률밀도함수(cdf) $F(x \mid \mu, \beta) = \frac{1}{[1+\exp(-(x-\mu)/\beta)]}$

$\beta_1 > 0$ 인 경우 logistic 회귀 모형 $\pi(x) = \exp(\beta_0 + \beta_1 x)/[1+\exp(\beta_0 + \beta_1 x)]$ 은 Logistic 분포의 확률밀도함수와 유사하다.

Logistic regression 선은 만약 F 가 $\mu = 0, \tau = 1$ 인 표준 CDF 이면 $\pi(x) = F(\beta_0 + \beta_1 x)$ 형태를 갖는다. 이 선은 평균이 $-\alpha/\beta$ 이고 분산이 $\frac{\pi^2 \beta^2}{3}$ 인 Logistic 분포함수의 CDF 이다. 즉 Logit 변환은 logistic CDF 의 역함수이다.

Probit model

만약 F 가 표준화 정규분포 CDF Φ 이면 $\pi(x) = \Phi(\beta_0 + \beta_1 x)$ 는 Probit 모형이다. 정규분포의 꼬리가 Logistic 분포의 꼬리보다 얇으므로 $\pi(x)$ 가 더 빨리 0 이나 1 로 접근한다.

$$\mathrm{Probit}(\pi(x)) = \Phi^{-1}(\pi(x)) = \beta_0 + \beta_1 x$$

3.2.4. 모형 추정

Linear probability model: $\hat{\pi}(x) = \beta_0 + \beta_1 x$

Logit model: $\ln(\frac{\hat{\pi}(x)}{1-\hat{\pi}(x)}) = \beta_0 + \beta_1 x$

Probit model: $\Phi^{-1}(\hat{\pi}(x)) = \beta_0 + \beta_1 x$

Example Thymidine 주사 후 셀의 증식 활동 지수(LI)와 암 환자 고통 완화 연구: 고통이 완화된 것을 성공이라고 간주하고 이를 1 로 코딩 하였다. LI 의 14 수준에서 27 환자들이 (관측치) 조사되었다.

LI	총 환자 수	고통 완화 환자 수	$\pi(x)$ 관측치
8	2	0	0
10	2	0	0
12	3	0	0
14	3	0	0
16	3	0	0
18	1	1	1
20	3	2	2/3
22	2	1	1/2
24	1	0	0
26	1	1	1
28	1	1	1
32	1	0	0
34	1	1	1
38	3	2	2/3

```
DATA CANCER;
       INPUT LI CASE GOOD @@;
       CARDS;
8      2      0      10     2      0      12     3      0
14     3      0      16     3      0      18     1      1
20     3      2      22     2      1      24     1      0
26     1      1      28     1      1      32     1      0
34     1      1      38     3      2
;
RUN;
TITLE 'Linear Link Function';
PROC GENMOD DATA=CANCER;
       MODEL GOOD/CASE=LI /LINK=IDENTITY DIST=NORMAL;
       OUTPUT OUT=OUT1 PRED=YHAT_LI;
RUN;
TITLE 'Logit Link Function';
PROC GENMOD DATA=CANCER;
       MODEL GOOD/CASE=LI /LINK=LOGIT DIST=BIN;
       OUTPUT OUT=OUT2 PRED=YHAT_LO;
RUN;
TITLE 'Probit Link Function';
PROC GENMOD DATA=CANCER;
```

```
        MODEL GOOD/CASE=LI /LINK=PROBIT;
        OUTPUT OUT=OUT3 PRED=YHAT_PR;
RUN;
DATA FIN;
        MERGE OUT1 OUT2 OUT3;
RUN;
PROC PRINT DATA=FIN;RUN;
```

프로그램 설명

- GENMOD 는 GENeralized linear MODel 의 약어이다.
- Model 은 모형을 설정한다. 반응변수=종속변수들 형태를 갖춘다.
- LINK 는 연결함수를 지정한다.
  - ✓ Linear Probability Model 에서는 반응변수는 정규분포 함수 설정 (DIST=Normal)
  - ✓ Logit model 에서는 반응변수 분포를 이항분포로 설정
  - ✓ Probit 에서는 DIST 가 필요 없다.
- OUTPUT 문(statement)은 모형 추정 결과를 저장한다.
- OUT 옵션은 결과를 저장하는 SAS data 이름을 지정한다.
- 어떤 추정 결과를 저장할지 지정한다.
  - ✓ PRED=YHAT1 는 예측치(predicted value)를 YHAT1 변수명에 저장한다. P=
  - ✓ RES= / U= / L=

Data Set	WORK.CANCER
Distribution	Normal
Link Function	Identity
Response Variable (Events)	GOOD
Response Variable (Trials)	CASE
Observations Used	14
Number Of Events	9
Number Of Trials	27

Linear Prob. Model

$\hat{\pi}(x) = -0.2507 + 0.0288 * LI$

$-2(L_2 - L_1)$

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	12	4.8145	0.4012
Scaled Deviance	12	54.2159	4.5180
Pearson Chi-Square	12	2.4072	0.2006
Scaled Pearson X2	12	27.1080	2.2590
Log Likelihood		-11.3542	

경고: The relative Hessian convergence criterion of 0.22589714 than the limit of 0.0001. The convergence is questionable.

설명변수 LI 는 유의하다.
양의 부호(0.0288)를
가지므로 LI 가
증가할수록 병 완화
비율은 높아진다

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	-0.2507	0.1380	-0.5212	0.0198	3.30	0.0693
LI	1	0.0288	0.0063	0.0166	0.0411	21.28	<.0001
Scale	1	0.2980	0.0364	0.2346	0.3785		

노트: The scale parameter was estimated by maximum likelihood.

Data Set	WORK.CANCER
Distribution	Binomial
Link Function	Logit
Response Variable (Events)	GOOD
Response Variable (Trials)	CASE
Observations Used	14
Number Of Events	9
Number Of Trials	27

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	12	15.6622	1.3052
Scaled Deviance	12	15.6622	1.3052
Pearson Chi-Square	12	13.3333	1.1111
Scaled Pearson X2	12	13.3333	1.1111
Log Likelihood		-13.0365	

Algorithm converged.

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	-3.7771	1.3786	-6.4792	-1.0751	7.51	0.0061
LI	1	0.1449	0.0593	0.0286	0.2612	5.96	0.0146
Scale	0	1.0000	0.0000	1.0000	1.0000		

Logit Model

$$\ln\left(\frac{\hat{\pi}(x)}{1-\hat{\pi}(x)}\right) = -3.771 + 0.1449 * LI$$

$$\hat{\pi}(x) = \frac{1}{1+\exp-(-3.771+0.1449*LI)}$$

설명변수 LI 는 유의하다.
양의 부호(0.1449)를
가지므로 위의 식에서
LI 가 증가할수록
병 완화 비율 높아진다

Data Set	WORK.CANCER
Distribution	Binomial
Link Function	Probit
Response Variable (Events)	GOOD
Response Variable (Trials)	CASE
Observations Used	14
Number Of Events	9
Number Of Trials	27

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	12	15.4437	1.2870
Scaled Deviance	12	15.4437	1.2870
Pearson Chi-Square	12	13.2661	1.1055
Scaled Pearson X2	12	13.2661	1.1055
Log Likelihood		-12.9272	

Algorithm converged.

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	-2.3178	0.7795	-3.8457	-0.7899	8.84	0.0029
LI	1	0.0878	0.0328	0.0236	0.1521	7.19	0.0073
Scale	0	1.0000	0.0000	1.0000	1.0000		

Probit Model

$$\Phi^{-1}(\hat{\pi}(x)) = -2.3178 + 0.0878 * LI$$

$$\hat{\pi}(x) = \Phi(-2.3178 + 0.0878 * LI)$$

설명변수 LI 는 유의하다.
양의 부호(0.0328)를
가지므로 위의 식에서
LI 가 증가할수록
병 완화 비율 높아진다

Obs	LI	CASE	GOOD	YHAT_LI	YHAT_LO	YHAT_PR
1	8	2	0	-0.01999	0.06797	0.05316
2	10	2	0	0.03769	0.08879	0.07504
3	12	3	0	0.09537	0.11519	0.10319
4	14	3	0	0.15305	0.14817	0.13832
5	16	3	0	0.21074	0.18857	0.18084
6	18	1	1	0.26842	0.23693	0.23072
7	20	3	2	0.32610	0.29320	0.28747
8	22	2	1	0.38378	0.35660	0.35009
9	24	1	0	0.44146	0.42545	0.41707
10	26	1	1	0.49914	0.49733	0.48656
11	28	1	1	0.55683	0.56931	0.55646
12	32	1	0	0.67219	0.70234	0.68914
13	34	1	1	0.72987	0.75918	0.74829
14	38	3	2	0.84523	0.84911	0.84626

$\hat{\pi}(x)$

그래프 그리기

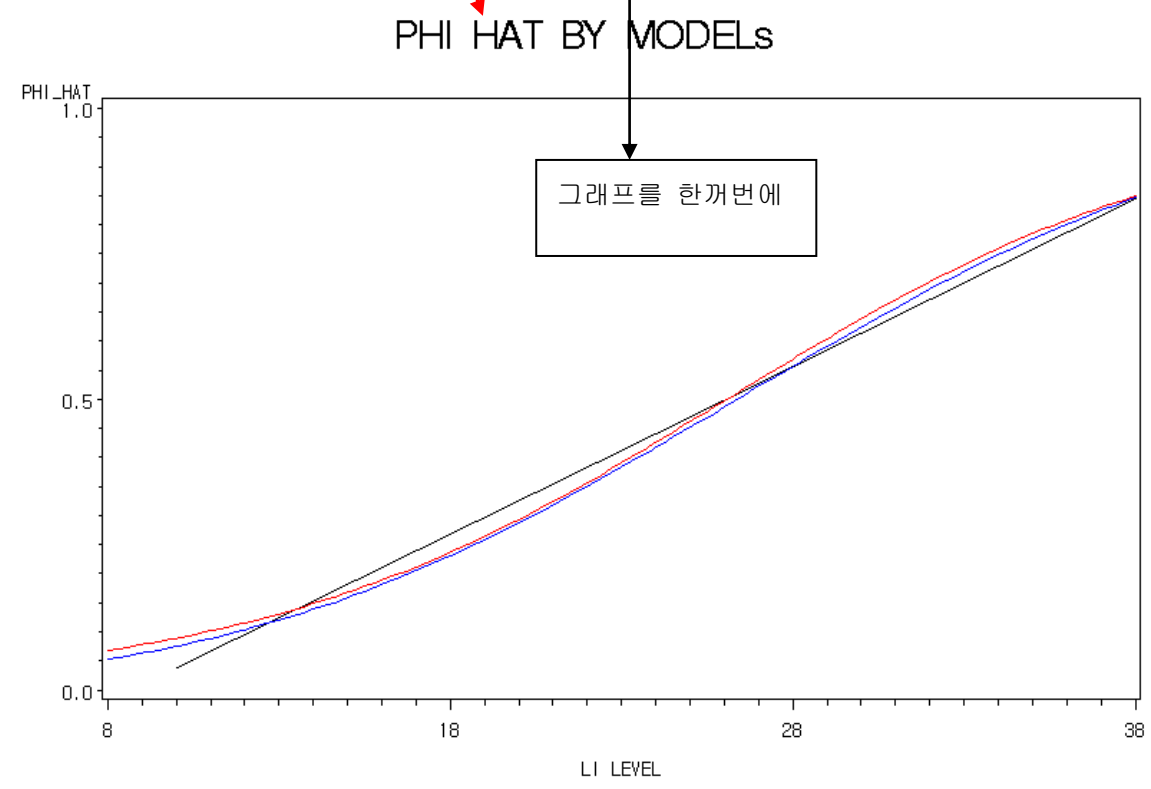
```
SYMBOL1 I=L3 V=NONE C=BLACK;
SYMBOL2 I=L3 V=NONE C=RED;
SYMBOL3 I=L3 V=NONE C=BLUE;
AXIS1 ORDER=0 TO 1 BY 0.5
        LABEL=('PHI_HAT');
AXIS2 ORDER=8 TO 38 BY 10
        LABEL=('LI LEVEL');
TITLE 'PHI HAT BY MODELs';
PROC GPLOT DATA=FIN;
PLOT (YHAT_LI YHAT_LO YHAT_PR)*LI /OVERLAY VAXIS=AXIS1 HAXIS=AXIS2;
RUN;
```

Symbol: 선들에 대한 옵션

V= value C=color I=interpolate

Axis 는 축에 관한 옵션

ORDER= 눈금, LABEL=축 이름



그래프를 한꺼번에

Recall: 회귀분석에서의 접근 [1999년 1학기 강의]

- 종속변수가 분류형 변수인 경우 설명변수와 인과 관계를 살펴보는데 사용된다.
- 종속변수의 수준이 3 개 이상인 경우 LOGISTIC 모형을 사용하는 것이 아니라 CATMOD 를 사용해야 한다고 언급한 책이 있다. 그러나 CATMOD 는 CATegorical data MODeling 의 약어로 분류변수 자료 모형화이며, LOGISTIC 모형은 CATMOD 기법의 한 부분입니다.
- LOGISTIC 모형은 종속 변수의 수준이 3 개 이상인 경우에도 가능하나 종속변수가 분류형 변수 중 ordinal(순서형⇔명목형: nominal)일 경우만 가능하다. 순서형 변수? 기업의 크기 (대, 중,소), 건강 상태 (양호, 보통, 불량), 학점(A, B, C, D) 등 크기 순서에 의해 분류된 경우

ODDs 개념 (Betting 의 기준이 된다)

- p/(1-p): 어떤 사건이 발생할 가능성[p=0.5 일 경우 1 이다. 기준]
- 한국이 2002 년 16 강에 들어갈 확률 0.1 이면 1/9 이 Odds 이다. => 1$ betting, 9$ return
- 브라질이 2002 년 16 강에 들어갈 확률 0.8 이면 4 가 Odds 이다. => 4$ betting, 1$ return

OLS 방법의 단점

- 결정계수가 매우 낮고 회귀 계수의 t-검정통계량 값이 맞다.
- E(Y)=Prob(Y=발생)이므로 해석의 어려움이 있다. 실제 값은 0, 1, 2 이나 예측치는 그 값이 발생할 확률이다.
- 분류형 변수 특성 상 이분산의 가능성이 높다.

ODDS TRANSFORMATION

- p/(1-p)를 odds transformation 이라 한다.
- p=Pr(Y=1)일 확률이라 하자. p 는 0 과 1 사이이므로 odds 는 0 과 ∞이다.
- Log 변환을 하면 LOG{p/(1-p)}는 -∞와 ∞사이의 값이므로

$$LOGIT(p) = \ln(\frac{p}{1-p}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + .. + \beta_p X_p + e$$

$$\Rightarrow \quad p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + .. + \beta_p X_p + e)}}$$

모형의 적합성 검정 및 회귀계수 유의성 검정

- –2Log L , AIC(Akaike Information Criterion) Schwartz Criterion=> Adjusted 결정계수와 유사한 개념
- 회귀계수의 유의성 검정은 Wald 의 Chi-square 검정통계량을 이용한다.

The LOGISTIC Procedure

```
PROC LOGISTIC DATA=LOGIT;
   MODEL Y=X1-X5/CTABLE INFLUENCE;
   OUTPUT OUT=OUT1 P=YHAT;
RUN;
PROC PRINT DATA=OUT1;
RUN;
```

Response Profile

Ordered Value	Y	Count
1	0	33
2	1	32

Event

No Event

Model Fitting Information and Testing Global Null Hypothesis BETA=0

Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
AIC	92.094	13.541	.
SC	94.268	26.587	.
-2 LOG L	90.094	1.541	88.553 with 5 DF (p=0.0001)
Score	.	.	41.640 with 5 DF (p=0.0001)

모형의 유의성 검정 => 모든 회귀계수는 0 이다.
P-값이 0.0001 이므로 귀무가설 기각

Analysis of Maximum Likelihood Estimates

표준화 회귀계수

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
INTERCPT	1	18.4986	15.6692	1.3937	0.2378	.	.
X1	1	-0.3601	0.4981	0.5229	0.4696	-8.229764	0.698
X2	1	-0.3064	0.2870	1.1397	0.2857	-12.092019	0.736
X3	1	-0.3442	0.3233	1.1332	0.2871	-8.376680	0.709
X4	1	0.00200	0.0624	0.0010	0.9744	0.205829	1.002
X5	1	-5.7610	5.9634	0.9333	0.3340	-3.416152	0.003

Association of Predicted Probabilities and Observed Responses

Concordant =100.0%	Somers' D = 1.000
Discordant = 0.0%	Gamma = 1.000
Tied = 0.0%	Tau-a = 0.508
(1056 pairs)	c = 1.000

설명변수 유의성 검정 => 회귀계수는 0이다.
P-값이 0.05 이하인 설명변수만 유의

Classification Table

	Correct		Incorrect		Percentages				
Prob Level	Event	Non-Event	Event	Non-Event	Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
0.000	33	0	32	0	50.8	100.0	0.0	49.2	.
0.020	31	29	3	2	92.3	93.9	90.6	8.8	6.5
0.040	31	29	3	2	92.3	93.9	90.6	8.8	6.5
[생략]									
0.500	31	29	3	2	92.3	93.9	90.6	8.8	6.5
0.520	31	30	2	2	93.8	93.9	93.8	6.1	6.3
0.540	31	30	2	2	93.8	93.9	93.8	6.1	6.3
0.560	31	30	2	2	93.8	93.9	93.8	6.1	6.3
0.580	31	30	2	2	93.8	93.9	93.8	6.1	6.3
[생략]									
0.740	31	30	2	2	93.8	93.9	93.8	6.1	6.3
0.760	31	30	2	2	93.8	93.9	93.8	6.1	6.3
0.780	31	30	2	2	93.8	93.9	93.8	6.1	6.3
0.800	30	30	2	3	92.3	90.9	93.8	6.3	9.1
0.820	30	30	2	3	92.3	90.9	93.8	6.3	9.1
0.840	30	30	2	3	92.3	90.9	93.8	6.3	9.1
[생략]									
0.980	30	30	2	3	92.3	90.9	93.8	6.3	9.1
1.000	0	32	0	33	49.2	0.0	100.0	.	50.8

Sensitivity=Event 반응 중 Event로 예측된 비율

Specificity=No event 중 No Event로 예측된 비율

어떤 Phat값을 기준으로 반응변수(종속변수)를 분류할지 결정해야 한다.

결정은 분석자의 주관에 의한다. 일반적으로 0.5를 기준으로 하면 무방하나, Classification Table의 정보를 이용해 오분류(misclassification) cost가 가장 적은 영역의 Phat를 이용하는 것이 바람직하다.

이 예제에서는 나는 0.6을 선택했다.

영향치나 이상치를 발견하는 검정통계량이다.

C, Cbar는 Cook distance에 근거를 두고 있다.

DIFDEV, DIFCHISQ는 ill-fitted 관측치를 발견하는 사용된다.

The LOGISTIC Procedure
WARNING: The validity of the model fit is questionable.

Regression Diagnostics

Case Number	Covariates X1	X2	X3	X4	X5	Pearson Residual Value	(1 unit = 0.06) -8 -4 0 2 4 6 8
1	36.7000	-62.8000	-89.5000	54.1000	1.7000	0	*
2	24.0000	3.3000	-3.5000	20.9000	1.1000	0.1531	| *
3	-61.6000	-120.8	-103.2	24.7000	2.5000	0	*
4	-1.0000	-18.1000	-28.8000	36.2000	1.1000	8.102E-7	*
5	18.9000	-3.8000	-50.6000	26.4000	0.9000	3.477E-6	*
6	-57.2000	-61.2000	-56.2000	11.0000	1.7000	0	*
7	3.0000	-20.3000	-17.4000	8.0000	1.0000	6.518E-6	*
8	-5.1000	-194.5	-25.8000	6.5000	0.5000	0	*
9	17.9000	20.8000	-4.3000	22.6000	1.0000	0.4861	| *
10	5.4000	-106.1	-22.9000	23.8000	1.5000	0	*

[생략]

Case Number	Deviance Residual Value	(1 unit = 0.08) -8 -4 0 2 4 6 8	Hat Matrix Diagonal Value	(1 unit = 0.06) 0 2 4 6 8 12 16	INTERCPT Dfbeta Value	(1 unit = 6.14) -8 -4 0 2 4 6 8
1	0	*	1.38E-17	*	0	*
2	0.2153	| *	0.7022	*	0.1268	*
3	0	*	6.13E-40	*	0	*
4	1.146E-6	*	2.8E-10	*	1.22E-11	*
5	4.917E-6	*	3.749E-9	*	1.59E-10	*
6	0	*	1.52E-26	*	0	*
7	9.218E-6	*	1.186E-8	*	6.64E-10	*
8	0	*	2.08E-34	*	0	*
9	0.6513	*	0.9824	*	14.9499	| *
10	0	*	1.3E-18	*	0	*

[생략]

Case Number	X1 Dfbeta Value	(1 unit = 5.9) -8 -4 0 2 4 6 8	X2 Dfbeta Value	(1 unit = 4.15) -8 -4 0 2 4 6 8	X3 Dfbeta Value	(1 unit = 2.87) -8 -4 0 2 4 6 8
1	0	*	0	*	0	*
2	0.1823	*	-0.3625	*	0.1328	*
3	0	*	0	*	0	*
4	-111E-13	*	-658E-14	*	-107E-13	*
5	-153E-12	*	-857E-13	*	-191E-12	*
6	0	*	0	*	0	*
7	-466E-12	*	-492E-12	*	-446E-12	*
8	0	*	0	*	0	*
9	-15.8791	*	0.6392	*	-12.3060	*
10	0	*	0	*	0	*

[생략]

Case Number	X4 Dfbeta Value	X5 Dfbeta Value	C

Case Number	(1 unit = 3.51) Value	-8 -4 0 2 4 6 8	Value	(1 unit = 5.66) -8 -4 0 2 4 6 8	Value	(1 unit = 489) 0 2 4 6 8 12 16
1	0	| * |	0	| * |	0	|* |
2	-0.2647	| * |	-0.1391	| * |	0.1857	|* |
3	0	| * |	0	| * |	0	|* |
4	5.99E-12	| * |	-981E-14	| * |	1.84E-22	|* |
5	5.81E-11	| * |	-114E-12	| * |	4.53E-20	|* |
6	0	| * |	0	| * |	0	|* |
7	1.93E-10	| * |	-558E-12	| * |	5.04E-19	|* |
8	0	| * |	0	| * |	0	|* |
9	6.7576	| | * |	-9.0523	| * |	752.6	| * |
10	0	| * |	0	| * |	0	|* |
		CBAR		DIFDEV		DIFCHISQ

Case Number	(1 unit = 2.7) Value	0 2 4 6 8 12 16	Value	(1 unit = 2.73) 0 2 4 6 8 12 16	Value	(1 unit = 2.72) 0 2 4 6 8 12 16
1	0	|* |	0	|* |	0	|* |
2	0.0553	|* |	0.1016	|* |	0.0787	|* |
3	0	|* |	0	|* |	0	|* |
4	1.84E-22	|* |	1.31E-12	|* |	6.56E-13	|* |
5	4.53E-20	|* |	2.42E-11	|* |	1.21E-11	|* |
6	0	|* |	0	|* |	0	|* |
7	5.04E-19	|* |	8.5E-11	|* |	4.25E-11	|* |
8	0	|* |	0	|* |	0	|* |
9	13.2176	| * |	13.6419	| * |	13.4539	| * |
10	0	|* |	0	|* |	0	|* |

OBS	X1	X2	X3	X4	X5	Y	_LEVEL_	YHAT
1	36.7	-62.8	-89.5	54.1	1.7	0	0	1.00000
2	24.0	3.3	-3.5	20.9	1.1	0	0	0.97710
3	-61.6	-120.8	-103.2	24.7	2.5	0	0	1.00000
4	-1.0	-18.1	-28.8	36.2	1.1	0	0	1.00000
5	18.9	-3.8	-50.6	26.4	0.9	0	0	1.00000
6	-57.2	-61.2	-56.2	11.0	1.7	0	0	1.00000
7	3.0	-20.3	-17.4	8.0	1.0	0	0	1.00000
8	-5.1	-194.5	-25.8	6.5	0.5	0	0	1.00000
9	17.9	20.8	-4.3	22.6	1.0	0	0	0.80888
10	5.4	-106.1	-22.9	23.8	1.5	0	0	1.00000
[생략]								
64	60.3	59.5	7.0	226.6	2.0	1	0	0.00000
65	17.9	16.3	20.4	105.6	1.0	1	0	0.00402
66	24.7	21.7	-7.8	118.6	1.6	.	0	0.03417

Yhat는 Pr(Y=Event)의 예측치이므로 Yhat의 값이 1에 가까우면 그 관측치는 Event(Y=1)로 분류된다.

앞에서는 분류 기준을 0.6으로 설정하였으므로 Yhat의 값이 0.6보다 크면 관측치를 1그룹(성공; event)으로 분류하고 0.6 이하이면 0 그룹(실패; non-event)으로 분류한다.

66번째 관측치는 실패 그룹으로 분류된다.

유의한 설명변수를 찾는 방법[Selection of Variables]
일반 회귀모형과 동일하게 변수 선택을 할 수 있다.
[option도 동일하다]

```
PROC LOGISTIC DATA=LOGIT;
   MODEL Y=X1-X5/SELECTION=STEPWISE SLENTRY=0.05;
RUN;
```

NOTE: Model building terminates because the last variable entered is removed by the Wald statistic criterion.

Summary of Stepwise Procedure

Step	Variable Entered	Variable Removed	Number In	Score Chi-Square	Wald Chi-Square	Pr > Chi-Square
1	X2		1	31.0487	.	0.0001
2	X3		2	4.7115	.	0.0300
3		X3	1	.	2.8334	0.0923

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
INTERCPT	1	1.1717	0.8103	2.0908	0.1482	.	.
X2	1	-0.1738	0.0568	9.3800	0.0022	-6.859194	0.840

```
PROC LOGISTIC DATA=LOGIT;
   MODEL Y=X2/CTABLE INFLUENCE;   OUTPUT OUT=OUT1 P=YHAT;
RUN;
PROC PRINT DATA=OUT1;
```

Classification Table

Prob Level	Correct Event	Correct Non-Event	Incorrect Event	Incorrect Non-Event	Percentages Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
[생략]									
0.300	32	30	2	1	95.4	97.0	93.8	5.9	3.2
0.320	32	30	2	1	95.4	97.0	93.8	5.9	3.2
0.340	32	30	2	1	95.4	97.0	93.8	5.9	3.2
0.360	32	30	2	1	95.4	97.0	93.8	5.9	3.2
0.380	32	30	2	1	95.4	97.0	93.8	5.9	3.2
0.400	32	30	2	1	95.4	97.0	93.8	5.9	3.2
0.420	31	30	2	2	93.8	93.9	93.8	6.1	6.3
0.440	31	30	2	2	93.8	93.9	93.8	6.1	6.3
0.460	31	30	2	2	93.8	93.9	93.8	6.1	6.3
0.480	31	31	1	2	95.4	93.9	96.9	3.1	6.1
0.500	31	31	1	2	95.4	93.9	96.9	3.1	6.1
0.520	31	31	1	2	95.4	93.9	96.9	3.1	6.1
0.540	31	31	1	2	95.4	93.9	96.9	3.1	6.1
0.560	31	31	1	2	95.4	93.9	96.9	3.1	6.1
0.580	31	31	1	2	95.4	93.9	96.9	3.1	6.1

0.600	30	31	1	3	93.8	90.9	96.9	3.2	8.8
0.620	30	31	1	3	93.8	90.9	96.9	3.2	8.8

[생략]

OBS	X1	X2	X3	X4	X5	Y	_LEVEL_	YHAT
1	36.7	-62.8	-89.5	54.1	1.7	0	0	0.99999
2	24.0	3.3	-3.5	20.9	1.1	0	0	0.64523
3	-61.6	-120.8	-103.2	24.7	2.5	0	0	1.00000

[생략]

9	17.9	20.8	-4.3	22.6	1.0	0	0	0.07990
10	5.4	-106.1	-22.9	23.8	1.5	0	0	1.00000

[생략]

65	17.9	16.3	20.4	105.6	1.0	1	0	0.15956
66	24.7	21.7	-7.8	118.6	1.6	.	0	0.06913

9번째 관측치가 오분류. 그리고 대체로 Yhat의 값들이 0혹은 1로부터 멀어져 중앙값으로 쏠리는 경향이 있다(예; 관측치 2).

그러나 오분류 비율은 이전 모형에 비해 감소하였다. 적은 설명변수로 분류의 효율을 높였다.

만약 0.5 에 의해 집단을 분류하려면 다음 프로그램 이용...

```
DATA FIN;
    SET OUT1;
    IF (YHAT>0.5) THEN GROUP='NON-EVENT';
    IF (YHAT<=0.5) THEN GROUP='EVENT';
RUN;
PROC PRINT DATA=FIN; RUN;
```

Recall: 다변량 분석 측면 [2001년 1학기]

판별 분석(DA)은 판별 변수가 모두 측정형 (연속형: continuous, measurement, metric))인 경우 사용할 수 있다. 물론 decision tree 방법(CART, CHAID)인 경우 판별 변수가 이산형이나 순서형 분류형 변수인 경우도 가능하지만...

로지스틱 회귀 분석 (logistic regression)은 혹은 Logit(로짓) 회귀 분석과 동일하고 차이가 있다면 종속변수가 (독립변수) 이진(binary: 가질 수 있는 값이 실패/성공, 정품/불량 등과 같이 가질 수 있는 값이 2 개인 경우)인 경우 분석하는 것이다. 일반 회귀 분석은 종속변수가 연속형이어야 한다.

로지스틱 회귀 분석에서 종속 변수 값은 0, 1(사건: 성공, 불량)로 입력된다. 칠면조 예제를 생각해 보자. 야생 칠면조는 경우 y=1, 사육 칠면조는 y=0 으로 하여 회귀 분석하면 된다. 로지스틱 회귀 분석에서는 종속변수가 1 혹은 0 을 가질 확률을 추정하게 된다. 그러므로 이 확률을 이용하여 개체를 분류할 수 있다. 즉 어떤 개체에 대해 Pr(y=1)의 추정치가 0.5 보다 크면 야생 칠면조로 0.5 보다 작으면 사육 칠면조로 분류하면 된다.

로지스틱 회귀분석은 이진형 반응변수 뿐 아니라 반응변수가 순서형(ordinal) 분류형인 경우 사용할 수 있습니다. 예를 들면 종속 변수가 고객의 신용도이고 이 변수가 (상, 중, 하) 분류되어 있는 경우 사용할 수 있습니다. 종속변수의 수준이 3 개 이상인 경우 LOGISTIC 모형을 사용하는 것이 아니라 CATMOD 를 사용해야 한다고 언급한 책이 있다. 그러나 CATMD 는 CATegorical data MODeling 의 약어로 분류변수 자료 모형화이며, LOGISTIC 모형은 CATMOD 기법의 한 부분입니다

일반 선형 회귀 모형

$$y_i = f(\underline{x}) = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + .. + \beta_p x_{pi} + e_i, \quad e_i \sim iidN(0, \sigma^2)$$

로지스틱 회귀모형의 종속 변수는 0 과 1 두 값만 가지므로(더 이상 정규분포를 따르지 않는다) 결정계수(R^2)가 매우 낮고 F-검정이나 t-검정을 사용하여 모형, 회귀 계수 추정을 할 수 없다.

ODDS & ODDS transformation

p/(1-p): 어떤 사건이 발생할 가능성[p=0.5 일 경우 1 이다. 기준]

한국이 2002 년 16 강에 들어갈 확률 0.1 이면 1/9 이 Odds 이다. => 1$ betting, 9$ return

브라질이 2002 년 16 강에 들어갈 확률 0.8 이면 4 가 Odds 이다. => 4$ betting, 1$ return

odds transformation: p*=p/(1-p)

로지스틱 회귀 모형

종속 변수를 $p_i = \Pr(Y=1)$ 라고 생각해 보면 종속 변수는 어떤 사건이 일어날 확률이 (Y=1) 된다. 그리고 여기에 ODDS 변환을 해 보자.

$$p_i^* = \frac{p_i}{1-p_i}$$

확률 p_i 가 (0,1) 사이의 값을 가지므로 p_i^* 는 (0, ∞) 값을 가진다. $\ln(p_i^*)$ 변환을 하면 이 변수는 (-∞,∞) 값을 가지므로 다음과 같은 모형을 생각해 볼 수 있다.

$$\ln(\frac{p_i}{1-p_i}) = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + .. + \beta_p x_{pi} + e_i : \text{ 로지스틱 모형}$$

위의 모형을 다시 쓰면 다음과 같다.

$$p_i = \Pr(Y=1|\underline{x}) = \frac{e^{\{\alpha+\beta_1 x_{1i}+\beta_2 x_{2i}+..+\beta_p x_{pi}\}}}{1+e^{\{\alpha+\beta_1 x_{1i}+\beta_2 x_{2i}+..+\beta_p x_{pi}\}}} + e_i$$

$$p_i = \Pr(Y=1|\underline{x}) = \frac{1}{1+e^{-\{\alpha+\beta_1 x_{1i}+\beta_2 x_{2i}+..+\beta_p x_{pi}\}}} + e_i$$

모형의 적합성 검정 및 회귀계수 유의성 검정

모형 전체의 유의성은 –2Log L , AIC(Akaike Information Criterion) Schwartz Criterion 을 이용하고 (Adjusted 결정계수와 유사한 개념) 회귀계수의 유의성 검정은 Wald 의 Chi-square 검정통계량을 이용한다.

칠면조 예제를 사용해 Logistic 회귀분석을 실시하자.

자료 읽기 & Logistic 분석 맛보기

```
DATA TURKEY;
    INFILE 'C:\TEMP\TURKEY.TXT';
    INPUT ID $ HUM RAD ULN FEMUR TIN CAR D3P COR SCA TYPE $;
RUN;

PROC LOGISTIC DATA=TURKEY;
    MODEL TYPE=HUM--SCA;
RUN;
```

결과 해석

Model Information

Data Set	WORK.TURKEY
Response Variable	TYPE
Number of Response Levels	2
Number of Observations	33
Link Function	Logit
Optimization Technique	Fisher's scoring

자료의 수가 33
사육 19 마리
야생 14

Response Profile

Ordered Value	TYPE	Total Frequency
1	DOMESTIC	19
2	WILD	14

Event=1: Pr(Y=1) 사육

[중간 생략]

모형의 유의성 검정
전체적으로는 유의

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	44.9383	9	<.0001
Score	25.5460	9	0.0024
Wald	0.8136	9	0.9998

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	1	59.1293	1322.1	0.0020	0.9643
HUM	1	0.4081	33.6496	0.0001	0.9903
RAD	1	-2.3390	34.4951	0.0046	0.9459
ULN	1	1.9542	7.4043	0.0697	0.7918
FEMUR	1	2.0172	18.0251	0.0125	0.9109
TIN	1	-2.7279	19.9102	0.0188	0.8910

설명 변수 각각에 대한 유의성 검정, 그나마 다소 유의해 보이는 변수들에 대한 유의성 검정 결과 출력: 유의한 변수가 보이지 않는다.

경고: The validity of the model fit is questionable.

매우 유의하지 않은 변수들의 유의성 검정 결과 출력

Analysis of Maximum Likelih

Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
CAR	1	0.2361	6.7775	0.0012	0.9722
D3P	1	0.4089	5.0668	0.0065	0.9357
COR	1	-2.4476	15.5680	0.0247	0.8751
SCA	1	-0.2290	19.7547	0.0001	0.9908

변수 선택 프로그램 & 결과 해석

```
PROC LOGISTIC DATA=TURKEY;
    MODEL TYPE=HUM--SCA / SELECTION=STEPWISE SLE=0.2 SLS=0.1;
RUN;
```

방법은 STEPWISE 방법이고 SLE=0.2(ENTRY) SLS=0.1(STAY)이다. 로지스틱에서는 이 정도 값을 사용하면 된다.

Summary of Stepwise Selection

Step	Effect Entered	Effect Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
1	TIN		1	1	21.6200	.	<.0001
2	FEMUR		1	2	5.6931	.	0.0170
3		FEMUR	1	1	.	1.8810	0.1702

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	1	73.3164	27.1245	7.3060	0.0069
TIN	1	-0.5027	0.1863	7.2855	0.0070

최종적으로 선택된 변수는 TIN 변수이다.

반드시 넣고 싶은 변수 포함하여 변수 선택

```
PROC LOGISTIC DATA=TURKEY;
    MODEL TYPE=ULN TIN HUM RAD  FEMUR CAR D3P COR SCA /
    SELECTION=STEPWISE SLE=0.2 SLS=0.1 INCLUDE=2;
RUN;
```

최종적으로 선택된 변수는 TIN 변수 하나만이지만 처음 9 개 변수 모두를 넣고 로지스틱 분석한 결과 ULN 이 가장 유의하였다(p-값=0.7918). 그래서 변수 선택을 하되 처음 2 개의 변수를 반드시 포함하라는 옵션이 INCLUDE=2 이다. 이 경우 MODEL 문에 포함하기 원하는 변수를 반드시 제일 앞에 써야 한다.

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	1	85.1576	46.6006	3.3394	0.0676
ULN	1	0.5030	0.3443	2.1349	0.1440
TIN	1	-1.0890	0.6030	3.2613	0.0709

최종적으로 선택된 변수는 TIN 변수와 ULN 이다. 이 두개 외에 다른 변수는 유의하지 않았다.

개체 판별하기

```
PROC LOGISTIC DATA=TURKEY;
     MODEL TYPE=HUM--SCA
     / SELECTION=STEPWISE SLE=0.2 SLS=0.1 CTABLE;
RUN;
```

CTABLE 은 Classification table 로 개체 분류를 위한 정보를 제공한다.

앞에서 EVENT 는 DOMESTIC(사육) 칠면조였다.

Event 를 Event 로 정분류
사육을 사육으로 분류

Non-Event 를 Non-Event 로 정분류
야생을 야생으로 분류

Classification Table

Prob Level	Correct Event	Correct Non-Event	Incorrect Event	Incorrect Non-Event	Percentages Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
0.000	19	0	14	0	57.6	100.0	0.0	42.4	.
0.020	19	5	9	0	72.7	100.0	35.7	32.1	0.0
0.040	19	6	8	0	75.8	100.0	42.9	29.6	0.0
0.060	19	7	7	0	78.8	100.0	50.0	26.9	0.0
0.080	19	10	4	0	87.9	100.0	71.4	17.4	0.0
0.100	19	10	4	0	87.9	100.0	71.4	17.4	0.0
0.120	19	10	4	0	87.9	100.0	71.4	17.4	0.0
0.140	19	11	3	0	90.9	100.0	78.6	13.6	0.0
0.160	19	11	3	0	90.9	100.0	78.6	13.6	0.0
0.180	19	11	3	0	90.9	100.0	78.6	13.6	0.0
0.200	19	11	3	0	90.9	100.0	78.6	13.6	0.0
0.220	19	11	3	0	90.9	100.0	78.6	13.6	0.0
0.240	19	11	3	0	90.9	100.0	78.6	13.6	0.0
0.260	19	11	3	0	90.9	100.0	78.6	13.6	0.0
0.280	18	11	3	1	87.9	94.7	78.6	14.3	8.3
0.300	18	11	3	1	87.9	94.7	78.6	14.3	8.3
0.320	18	11	3	1	87.9	94.7	78.6	14.3	8.3
0.340	18	11	3	1	87.9	94.7	78.6	14.3	8.3
0.360	18	11	3	1	87.9	94.7	78.6	14.3	8.3
0.380	18	11	3	1	87.9	94.7	78.6	14.3	8.3
0.400	18	11	3	1	87.9	94.7	78.6	14.3	8.3
0.420	17	12	2	2	87.9	89.5	85.7	10.5	14.3
0.440	17	12	2	2	87.9	89.5	85.7	10.5	14.3
0.460	17	12	2	2	87.9	89.5	85.7	10.5	14.3
0.480	17	12	2	2	87.9	89.5	85.7	10.5	14.3
0.500	17	12	2	2	87.9	89.5	85.7	10.5	14.3
0.520	17	12	2	2	87.9	89.5	85.7	10.5	14.3
0.540	17	12	2	2	87.9	89.5	85.7	10.5	14.3
0.560	16	12	2	3	84.8	84.2	85.7	11.1	20.0
0.580	16	12	2	3	84.8	84.2	85.7	11.1	20.0
0.600	16	12	2	3	84.8	84.2	85.7	11.1	20.0
0.620	16	12	2	3	84.8	84.2	85.7	11.1	20.0

non-EVENT(야생)를 EVENT(사육)로 오분류

EVENT(사육)를 non-EVENT(야생)으로 오분류

0.3 을 cut-off 값으로 하면 어떨지… 만약 Pr(Y=1)예측치 값이 0.3 보다 크면 Event(사육)으로 0.3 보다 크면 야생으로 분류하면 된다.

0.4 에서도 오분류는 동일하게 4 개이다. 차이가 있다면 event 를 Non-event 로 오분류 할 가능성이 높다는 것이다. 그러므로 cost 를 생각하여 cut-off 선택은 분석자 자유.

새로운 개체분류 하기

```
DATA Temp;
    HUM=150; RAD=150; ULN=150; FEMUR=150;
    TIN=150; CAR=150; D3P=150; COR=150; SCA=150;
output;
RUN;
DATA ALL;
    SET TURKEY TEMP;
RUN;
PROC LOGISTIC DATA=ALL ;
    MODEL TYPE=HUM--SCA
    / SELECTION=STEPWISE SLE=0.2 SLS=0.1 CTABLE;
    OUTPUT OUT=OUTO P=PHAT;
RUN;

PROC PRINT DATA=OUTO;
RUN;
```

Obs	ID	HUM	RAD	ULN	FEMUR	TIN	CAR	D3P	COR	SCA	TYPE	_LEVEL_	PHAT
15	B791	.	132	148	138	145	775	.	106	128	WILD	DOMESTIC	.
16	B795	151	134	151	144	.	789	292	116	126	WILD	DOMESTIC	.
17	B819	158	135	151	146	152	790	289	111	125	WILD	DOMESTIC	0.04319
18	B081	.	135	149	.	149	789	.	111	123	WILD	DOMESTIC	.
19	B085	148	129	146	139	147	767	287	106	123	WILD	DOMESTIC	0.35791
20	B089	157	140	154	140	159	818	301	116	136	WILD	DOMESTIC	0.00134
21	B090	153	138	153	141	151	822	312	115	133	WILD	DOMESTIC	0.06944
22	B091	156	138	156	145	150	835	310	118	133	WILD	DOMESTIC	0.10981
23	B093	151	133	148	139	152	793	290	105	.	WILD	DOMESTIC	.
24	B097	153	135	150	144	158	772	276	102	123	WILD	DOMESTIC	0.00221
25	B099	152	140	151	144	158	792	303	111	122	WILD	DOMESTIC	0.00221
83		150	150	150	150	150	150	150	150	150		DOMESTIC	0.10981

LEVEL=에는 Event 의 수준을 나타낸다. 출력 결과를 보면 모두 DOMESTIC 이다. PHAT 는 Pr(y=1:event)의 추정치이므로 0.5 이상이면 Event 로 분류하고 그 미만이면 non-event 로 분류한다. 우리는 앞에서 0.3 을 cut-off 로 하였으므로 19 번째 개체는 DOMESTIC 으로 분류되어야 한다. 이것이 오분류이고 cut-OFF 가 0.3 인 경우 Event(사육)라고 잘못 분류할 2 개 중에 하나이다. (19 번째, 30 번째, 36 번째): Wild=>Domestic 으로 오분류

83 번째 새로운 개체는 Nonevent 인 Wild(야생)으로 분류한다. phat=0.1081

HOMEWORK #6-3 TAX.txt

TAX.txt 자료는 다음 변수에 대한 자료이다. 다음 절차에 의해 Logistic 분석을 실시하시오.

1) 적절한 변수를 선택하고 (유의수준=0.1)
2) 분석 결과를 해석하시오.
3) Classification Table 을 보고 적절한 Phat 기준을 선택하시오. (분류에 참고)

종속변수: PREP(세금 보고 전문가 이용=1, 자신이 직접=0)

독립변수: 1)MA (결혼 여부, 1=결혼, 0=미혼) Indicator 변수
2)SE (자기 사업=1, 취업=0) Indicator 변수
3)DEP (부양 가족 수) : 측정형 변수(연속형)
4)TR (세금 효율:rate) : 측정형 변수(연속형)
5)INCOME (소득) : 측정형 변수(연속형)

HOMEWORK #7

다음 자료는 혈압(X)에 따른 심장병 발병 확률(Y)의 차이가 있는지 알아보기 위하여 조사한 자료이다.

Blood Pressure		Heart Disease: Present	Heart Disease: Absent
<117	111.5	3	153
117–126	121,5	17	235
127–136	131.5	12	272
137–146	141.5	16	255
147–156	151.5	12	127
157–166	161.5	8	77
167–186	176.5	16	83
>186	191.5	8	35

Source: Reprinted with permission based on Cornfield (1962).

혈압이 구간으로 추정되어 있으므로 모형 적합 시에는 구간의 중앙값을 사용하시오.

1) Logit Model 을 적합(fit)하고 결과를 해석하시오.

2) Probit Model 을 적합(fit)하고 결과를 해석하시오.
3) 원 자료, Logit Model 예측치, Probit Model 예측치의 산점도 그래프를 하나의 그래프에 나타내시오.

3.3. Logit model for categorical explanatory variable

3.2.절에서는 설명 변수가 연속형(측정형)인 경우 Logit 모형을 살펴보았다. 여기서는 설명변수가 범주형(categorical)일 때 분석 방법을 다루기로 하겠다. 사실 설명변수가 범주형이고 link 함수가 Logit 이면 다음 장에 살펴볼 Log-linear 모형과 같지만 간단한 예제 형식으로 살펴보기로 하자.

3.3.1. Logit model for Ix2 table

반응변수 / 설명변수(X)	성공	실패	합계
1	$\pi_{y=1\|x=1}$ n_{11}	$\pi_{y=0\|x=1} = 1-\pi_{y=1\|x=1}$ n_{12}	n_{1+}
2	$\pi_{y=1\|x=2}$ n_{21}	$\pi_{y=0\|x=2} = 1-\pi_{y=1\|x=2}$ n_{22}	n_{2+}
i	$\pi_{y=1\|x=i}$ n_{i1}	$\pi_{y=0\|x=i} = 1-\pi_{y=1\|x=i}$ n_{i2}	n_{i+}
R	$\pi_{y=1\|x=r}$ n_{r1}	$\pi_{y=0\|x=r} = 1-\pi_{y=1\|x=r}$ n_{r2}	n_{r+}

예제 자료

Blood Pressure	Heart Disease: Present	Heart Disease: Absent	
<117	3	153	156
117–126	17	235	252
127–136	12	272	284
137–146	16	255	271
147–156	12	127	139
157–166	8	77	85
167–186	16	83	99
>186	8	35	43

Source: Reprinted with permission based on Cornfield (1962).

Logit model

$$\ln(\frac{\pi_{1|i}}{\pi_{2|i}}) = \alpha + \beta_i \quad \text{--- (1)}$$

모형에 대한 분석은 일원 분산 분석 (one-way ANOVA) 이랑 동일하다. 설명변수가 연속형이면 βx_i 로 회귀분석과 같다. 단지 종속 변수가 y_i 가 아니라 $\ln(\frac{\pi_{1|i}}{\pi_{2|i}})$ 이다.

β_i 는 행의 효과인데 이는 요인 효과와 동일하다. α 는 로짓의 평균이고 (종속 변수 평균) β_i 가 높을수록 i 행의 logit 값은 크고 $\pi_{y=1|x=i}$ 는 증가한다.

만약 각 행의 총 빈도가 고정이면 (n_{i+}) 반응 변수는 성공/실패만 있으므로 모수가 $\pi_{y=1|x=i}$ (설명 변수 수준인 i 인 경우, 즉 i 행의 성공 확률)인 Bernoulli 분포를 따른다. 그러므로 i 행의 성공 회수(n_{i1})는 Binomial $(n = n_{i+}, p = \pi_{y=1|x=i})$ 분포를 따른다. 그리고 행의 효과가 없다면(설명 변수의 요인 효과가 없다면, $\beta_1 = \beta_2 = ...\beta_r = 0$) Logit model 은 다음과 같이 줄어든다. $\beta_1 = \beta_2 = ...\beta_r = 0$ ←→ $\pi_{1|x=1} = \pi_{1|x=2} = ... = \pi_{1|x=r}$ 그러므로 독립성 검정이란 동일하다.

$$\ln(\frac{\pi_{1|i}}{\pi_{2|i}}) = \alpha$$

3.3.2. Logit models for higher dimension

범주형인 설명 변수가 2 개 이상이고 반응 변수가 2 진인 경우 Logit model 을 사용할 수 있다. 설명의 편리를 위하여 설명변수가 2 개(요인 A, 요인 B)일 때 알아보기로 하자. 두 설명 변수 수준을 각각 I, J 라 하자. 그러면 $\pi_{y=1|ij} + \pi_{y=0|ij} = 1$ 이다. 그러므로 분할표는 IxJx2 형태이고 Logit model 은

$$\ln(\frac{\pi_{1|ij}}{\pi_{2|ij}}) = \alpha + \beta_i + \gamma_j \quad \text{--- (2)}$$

이 모형에 대한 분석은 교차 항이 (interaction) 없는 이원(two-way) 분산분석과 동일하다. 만약 각 행의 총 빈도가 고정이면 (n_{ij+}) 반응 변수는 성공/실패만 있으므로 모수가 $\pi_{y=1|x=ij}$ (설명 변수 수준인 (i, j)인 Bernoulli 분포를 따른다. 그러므로 i 행의 성공 회수(n_{ij1})는 Binomial $(n = n_{i+}, p = \pi_{y=1|ij})$ 분포를 따른다. 한 설명변수(요인) A 의 (주)효과가 없다면($\beta_1 = \beta_2 = ...\beta_r = 0$) Logit model 은 다음과 같이 줄어든다. 물론 요인 B 의 주효과에 대해서도 같은 이론이 적용될 수 있다.

$$\ln(\frac{\pi_{1|i}}{\pi_{2|i}}) = \alpha + \beta_j$$

3.3.3. 예제: 범주형인 설명 변수가 하나이고 종속변수가 binary 인 경우

혈압에 따른 심장병 발병 비율의 차이가 있는지 알아보기 위하여 조사된 자료이다. [Cornfield (1962) Homework#7 자료와 동일] 혈압이 구간으로 측정되어 있으므로 이를 범주형 척도로 인지하자. Homework#7 에서는 Logit Regression Model 을 사용하려면 설명변수가 측정형이어야 하므로 구간의 중앙값으로 (111.5, 121.5, …, 176.5, 191.5) 사용하였다.

ML (Maximum Likelihood) 추정치

Saturated model (1)에 대하여 $\{\beta_i\}$의 선택 제약 조건에 상관없이 $\{\alpha+\beta_i\}$는 일정하고 이에 대한 ML 추정치는 표본 Logit 이다. 즉,

$$\hat{\alpha}+\hat{\beta}_1 = \log(3/153) = -3.93$$

Logit Regression Model

Logit Regression model 에 의해 모형을 추정하면

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	-6.0820	0.7243	-7.5017	-4.6624	70.51	<.0001
pressure	1	0.0243	0.0048	0.0148	0.0338	25.25	<.0001
Scale	0	1.0000	0.0000	1.0000	1.0000		

$$\ln\left(\frac{\pi_{y=1|x=i}}{\pi_{y=0|x=i}}\right) = -6.08 + 0.0243x_i$$

추정치

x_i (BP)	sample logit	obs(ML) $\hat{\pi}$	logit $\hat{\pi}$
111.5	-3.93183	0.01923	0.03330
121.5	-2.62637	0.06746	0.04209
131.5	-3.12090	0.04225	0.05307
141.5	-2.76867	0.05904	0.06672
151.5	-2.35928	0.08633	0.08357
161.5	-2.26436	0.09412	0.10420
176.5	-1.64625	0.16162	0.14352
191.5	-1.47591	0.18605	0.19446

1) Logit regression model 의 회귀 계수의 유의성 추정은 WALD 통계량에 의한다.

$$T = \frac{\hat{\beta}}{ASE(AsymStdErr)} \sim z(0,1) \text{ (SAS 에 출력)}$$

2) Pearson Chi-square 통계량

$$\chi^2 = \sum \frac{(n_{ij} - \hat{n}_{ij})}{\hat{n}_{ij}} \sim \chi^2(df = (r-1)(c-1)$$

$\hat{n}_{ij}$ 는 logit model 에 의해 추정된 기대 도수: (예) $\frac{e^{-1.42}}{1+e^{-1.42}} = 0.194$ ➔ 43*0.194=8.4

Table 4.4 Cross-Classification of Framingham Men by Blood Pressure and Heart Disease

	Heart Disease[a]	
Blood Pressure	Present	Absent
<117	3 (5.2)	153 (150.8)
117–126	17 (10.6)	235 (241.4)
127–136	12 (15.1)	272 (268.9)
137–146	16 (18.1)	255 (252.9)
147–156	12 (11.6)	127 (127.4)
157–166	8 (8.9)	77 (76.1)
167–186	16 (14.2)	83 (84.8)
>186	8 (8.4)	35 (34.6)

Source: Reprinted with permission based on Cornfield (1962).

분할표 검정에 의하면

3) Likelihood Ratio Test

$$G^2 = 2\sum\sum n_{ij} \log(\frac{n_{ij}}{\hat{n}_{ij}}) \sim \chi^2(df = (r-1)(c-1)$$

$\hat{n}_{ij}$ 는 logit model 에 의해 추정된 기대 도수로 1)와 동일하다.

LOG-LOG Link

$$\ln(-\ln(\pi(x)) = \alpha + \beta x$$

```
PROC GENMOD DATA=heart;
    MODEL present/total=pressure /LINK=LOG DIST=BIN;
    OUTPUT OUT=OUT2 PRED=yhat_lo;
RUN;
```

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	-5.8394	0.6820	-7.1762	-4.5026	73.30	<.0001
pressure	1	0.0221	0.0045	0.0132	0.0309	23.81	<.0001
Scale	0	1.0000	0.0000	1.0000	1.0000		