범주형 자료분석
Basics in Logit and Probit Models

김현우, PhD[1]

[1] 충북대학교 사회학과 조교수

March 23, 2025

진행 순서

1. 로지스틱 회귀분석의 원리
2. 회귀계수의 해석
3. 유의성 검정
4. 모형 적합도
5. 몇가지 주의사항

로지스틱 회귀분석의 원리

로지스틱 회귀분석의 원리

종속변수가 가변수라면 회귀분석이 어떻게 달라지는가 상상해보자.

- 단순히 생각하면 가변수를 그대로 종속변수로 한 회귀식에서 β_0와 β_1을 추정할 수 있다.
- 종속변수가 0 또는 1이므로 (이것을 마치 확률처럼 접근하면) "X가 한 단위 증가할 때 Y가 1이 될 확률(probability of being 1)은 β_1 만큼 증가한다"고 해석한다(Why?).

$$E(Y|X) = P(Y = 1|X) = \beta_0 + \beta_1 X$$

- 이런 타입의 회귀모형을 선형확률모형(linear probability model)이라고 부른다.
- 그러나 이것은 (1) $\widehat{Y}$가 0보다 작거나 1보다 큰 값이 나오기도 하고, (2) 회귀분석의 가정 가운데 등분산성(homoscedasticity)에도 위배된다(증명 생략).

로지스틱 회귀분석의 원리

- 본래 선형회귀식의 종속변수는 양적 변수이므로 당연히 $-\infty < E(Y|X) < \infty$ 를 전제한다.
- 그러나 가변수가 종속변수라면 반드시 $E(Y|X) = \{0, 1\}$ 여야만 한다.
- 가변수로 부호화(encoding)되었다는 것은 범주가 두 개라는 뜻이다.
- (가변수를 독립변수로 사용할 때와 마찬가지로) 어느 한 쪽이 기준집단(reference group) 또는 기저범주(baseline category)이 되어 분석에서 제외된다.
- 어느 쪽이 1이고 나머지가 0이 될지는 직접 결정한다.

로지스틱 회귀분석의 원리

가변수가 종속변수일 때는 종속변수를 살짝 바꾸면 된다.

- 먼저 오즈(odds) 또는 승산(勝算)은 다음과 같이 정의된다.

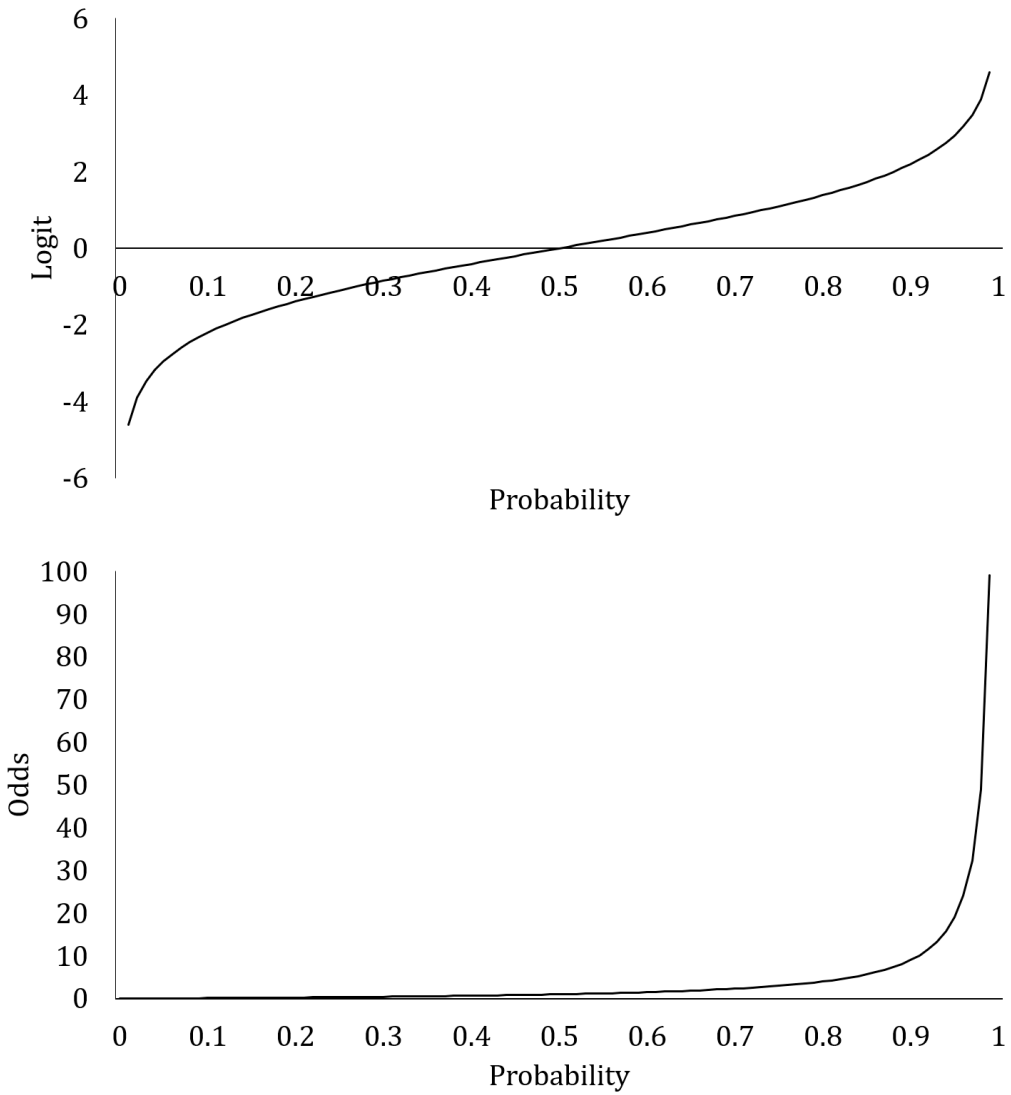
$$odds = \frac{P(Y=1|X)}{1-P(Y=1|X)}$$

- 오즈에 자연대수를 취한 값을 로그오즈 또는 로짓(logit)이라고 부른다.

$$logit = log_e \frac{P(Y=1|X)}{1-P(Y=1|X)} = ln \frac{P(Y=1|X)}{1-P(Y=1|X)}$$

- 로짓을 로지스틱 함수(logistic function)라고도 부른다.
- 확률은 정의상 [0, 1] 사이로 제한되지만, 로짓의 범위는 $[-\infty, \infty]$이므로 이제 $-\infty < E(Y|X) = P(Y=1|X) < \infty$ 가 성립한다.

로지스틱 회귀분석의 원리



로지스틱 회귀분석의 원리

- 이제 $E(Y|X) = \beta_0 + \beta_1 X$ 대신 다음의 회귀식을 사용한다.

$$ln \frac{P(Y=1|X)}{1-P(Y=1|X)} = \beta_0 + \beta_1 X$$

- 즉 종속변수로 $\{0, 1\}$인 원점수를 그대로 쓰지 않고, 범위 제한이 없는 로짓을 사용하는 편이 선형확률모형보다 낫다!
- 로짓을 종속변수로 하므로 로지스틱 회귀모형 또는 로짓(logit) 회귀모형이라고도 부른다.
- 구태여 덧붙이자면, (자료에서 관측되는) 종속변수가 $\{0, 1\}$이므로 이항(binary) 로지스틱(또는 이항 로짓) 회귀모형이라고도 부른다.
- 교과서에 따라서는 로지스틱 회귀모형과 로짓 회귀모형을 개념상 구분하지만, 실익이 크지 않으므로 너무 집착하지 않아도 된다.

회귀계수의 해석

회귀계수의 해석

로지스틱 회귀분석은 보통최소제곱이 아닌 다른 알고리즘을 요구한다.

- 선형회귀분석은 오차제곱합(sum of squared error)을 최소화하는 β_0와 β_1를 찾기 위해 보통최소제곱(OLS)이라는 알고리즘을 사용하였다.
- 로지스틱 회귀분석은 비선형모형(nonlinear model)인 탓에 OLS를 사용할 수 없다. 대신 최대우도법(maximum likelihood estimation; MLE)을 주로 사용한다.
- 최대우도법은 우도함수(likelihood function) λ를 정의한 뒤, 이것의 로그 우도(log-likelihood) $ln\lambda$를 극대화하는 β_0와 β_1를 찾는 알고리즘이다.

$$\underset{\beta_0,\beta_1}{\operatorname{argmax}}\, ln \prod_{i=1}^{n} f_i(Y_i) = \underset{\beta_0,\beta_1}{\operatorname{argmax}}\, ln\lambda(\beta_0, \beta_1; Y_i)$$

회귀계수의 해석

안타깝게도 로그오즈의 해석은 직관적이지 않다.

- 회귀식을 잘 살펴보면 로지스틱 회귀분석 특유의 β_1의 해석법을 깨달을 수 있다.

$$ln\frac{P(Y=1|X)}{1-P(Y=1|X)} = \beta_0 + \beta_1 X$$

- 가령 종속변수가 낙태권 지지(0=지지하지 않음; 1=지지함)이고 독립변수가 교육연수라고 하자.
- "교육연수 X가 한 단위 증가하면, 낙태권을 지지할 로그오즈는 β_1 만큼 증가한다."
- "로그오즈가 증가한다니... 이게 대체 무슨 소린가?" 나도 모르겠다.

회귀계수의 해석

그렇기 때문에 해석상 차라리 오즈비가 선호된다.

- 독립변수 X가 한 단위 더 증가한다면 예전 상태와 비교하여 오즈비(odds ratio; OR)는 다음과 같이 정의된다.

$$OR = \frac{odds_{X+1}}{odds_X} = \frac{\left(\frac{P(Y=1|X+1)}{1-P(Y=1|X+1)}\right)}{\left(\frac{P(Y=1|X)}{1-P(Y=1|X)}\right)}$$

$$= \frac{e^{\beta_0+\beta_1(X+1)}}{e^{\beta_0+\beta_1 X}} = e^{\beta_1(X+1)-\beta_1 X} = e^{\beta_1}$$

- 즉 e^{β_1}는 X가 $X+1$로 한 단위 증가할 때 '오즈가 달라지는 비율'을 보여준다 (Why?)!
- 그러므로 e^{β_1}를 그대로 오즈비라고 부른다.

회귀계수의 해석

- 오즈비는 정의상 비율(ratio)이므로, 가령 오즈비가 1.5라면 분모의 오즈보다 분자의 오즈가 50% 크다는 것을 뜻한다.
- 그러므로 '오즈가 달라지는 비율'을 다음과 같이 백분율로 나타낼 수 있다(Why?).

$$\Delta\% = 100 \cdot \left(e^{\beta_1} - 1\right)$$

- "X가 한 단위 증가하면, $Y = 1$의 오즈가 $100 \times (e^{\beta_1} - 1)$ 퍼센트 증가한다."
- 로지스틱 회귀계수의 오즈비 해석은 혼동하기 쉬우므로 많이 연습해야 한다.

회귀계수의 해석

예제 1. lbw.dta를 이용하여 저출생체중아 여부(low)를 종속변수로, 임신전 마지막 월경 당시 산모체중(lwt)을 독립변수로 하는 회귀식을 추정하시오. 두 변수의 연관성에 관해 해석하시오.

회귀계수의 해석

- 종속변수인 저출생체중아 여부(low)를 잘 살펴보면 0 (저출생체중 아님) 또는 1 (저출생체중)로 부호화되어 있으므로 (OLS를 사용한 선형회귀분석보다) 로지스틱 회귀분석이 바람직하다.
- Stata를 사용하여 다음과 같이 회귀식을 추정한다.

$$ln\frac{P(low=1|X)}{1-P(low=1|X)} = 0.996 - 0.014\text{lwt}$$

- "산모의 마지막 월경 당시 체중이 1 파운드 증가하면 산아가 저출생체중일 로그오즈는 0.014 만큼 감소한다."
- "산모의 마지막 월경 당시 체중이 1 파운드 증가하면 산아가 저출생체중일 오즈는 1.4% (=100 $\times$ $(e^{0.014}-1)$) 감소한다."

회귀계수의 해석

어쩌면 오즈비조차도 별로 직관적이지 않다.

- 오즈보다는 확률이야말로 우리에게 가장 직관적으로 와닿는 해석을 제공한다!
- 가령 "내가 이길 확률은 80% 정도 된다구" 라고 말하면 모를까, "내가 이길 오즈는 4 (=0.8/0.2) 정도 된다구" 라고 말하는 것은 몹시 이상하다.
- 그러므로 (로그)오즈를 아예 확률로 변환할 수 있다면 해석에 직관성을 더할 수 있다.
- 로그오즈와 확률의 관계는 다음과 같다(증명).

$$logit = ln\frac{P(Y=1|X)}{1-P(Y=1|X)} = \beta_0 + \beta_1 X$$

$$P(Y=1|X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}} = \frac{1}{1+e^{-(\beta_0+\beta_1 X)}}$$

- 그러면 추정된 $\widehat{\beta_0}$와 $\widehat{\beta_1}$를 대입하고 X의 증가에 따라 예측확률(predicted probability) $\widehat{P(Y=1|X)}$이 어떻게 변화하는지 그림을 그릴 수 있다.

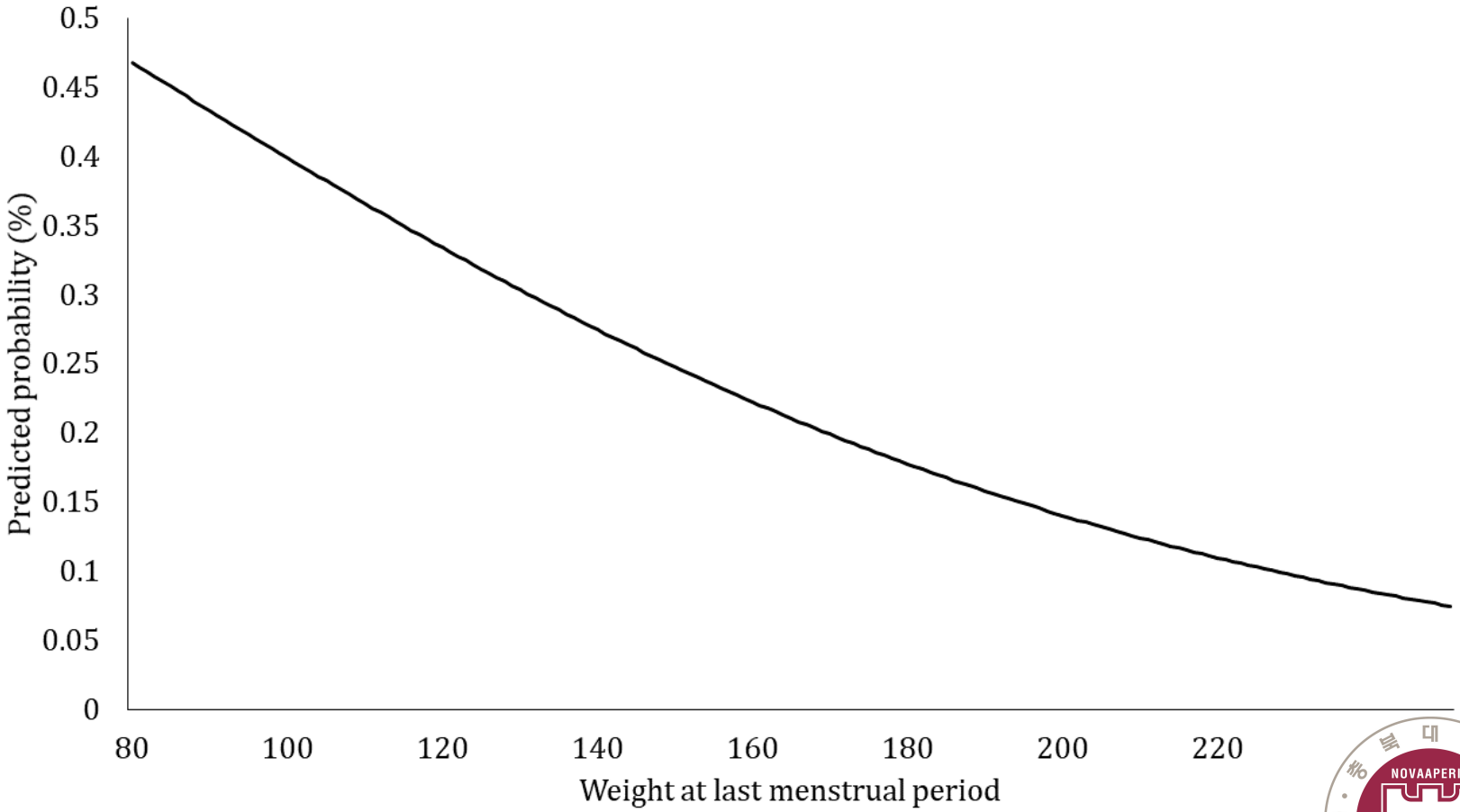
회귀계수의 해석

- Stata를 사용하여 예측확률 $P(\widehat{Y=1}|X)$을 계산할 수 있다. 엑셀에서도 직접 계산해보자(이것은 사실 훌륭한 연습이 된다).
- 임신전 체중(lwt)은 [80, 250] 파운드 사이에 놓이므로 각각의 예측확률을 계산한다.

	A	B	C	D	E	F
1			Variables in the Equation			
2			B	S.E.	Wald	df
3	Step 1[a]	Weight at last	-0.014	0.006	5.178	
4		Constant	0.996	0.785	1.608	
5	a. Variable(s) entered on step 1: Weight at last menstrual period.					
6						
7	X	b_0+b_1X	P			
8	80	=$C$4+$C$3*A8				
9	81	-0.14124	0.464748			
10	82	-0.15528	0.461258			
11	83	-0.16932	0.457771			

회귀계수의 해석

- 마지막으로 꺾은선 그래프(line chart)로 그린다.



유의성 검정

유의성 검정

로지스틱 회귀모형에서는 색다른 유의성 검정이 사용된다.

- OLS에서는 회귀계수의 통계적 유의성을 확인하기 위해 다음과 같은 t 검정(t test)을 사용하였다.

$$t = \frac{b - \beta}{SE_b} = \frac{b}{SE_b}$$

- t 분포의 꼬트머리를 그리고 그 면적을 통해 p 값을 구하는 방식으로 귀무가설 ($H_0 : \beta = 0$)을 기각할 수 있다.

유의성 검정

- 로지스틱 회귀모형과 같은 비선형모형에서는 자유도(degree of freedom; df)가 1인 χ^2 검정(χ^2 test)을 사용한다(증명 생략).

$$Wald = \left(\frac{b-\beta}{SE_b}\right)^2 = \left(\frac{b}{SE_b}\right)^2 \sim \chi_1^2$$

- 이 χ^2 검정을 고안한 Abraham Wald의 이름을 따 특별히 왈드 검정(Wald test)이라고 부른다.
- χ_1^2 값이 충분히 크면 통계적으로 유의하게 귀무가설($H_0 : \beta = 0$)을 기각할 수 있다.

유의성 검정

예제 2. lowbwt.sav는 저출생체중 여부(low) 뿐 아니라 산아의 체중 원점수(bwt)를 양적 변수로도 제공하고 있다(참고로 처출생체중아의 기준은 2,500g 이다). 모두 똑같이 smoke, age, lwt, ht를 독립변수로 하되, (1) 종속변수는 체중 원점수(bwt)인 선형회귀모형, (2) 종속변수는 저출생체중 여부(low)인 선형확률모형, 그리고 (3) 종속변수는 저출생체중 여부(low)인 로지스틱 회귀모형을 각각 추정하시오. 특히 임신중 흡연 여부에 초점을 두고 그 결과를 비교하시오.

유의성 검정

- 결과표를 엑셀로 옮겨 정리하자. 어떤 식으로 회귀식을 추정하던지 유의성 검정 결과는 거의 똑같다.

  (1) “(다른 변수의 영향력을 통제할 때) 산모가 임신중 흡연할 경우 산아의 체중은 그렇지 않은 산모보다 평균적으로 261.9g 적다.”

  (2) “(다른 변수의 영향력을 통제할 때) 산모가 임신중 흡연할 경우 산아가 저출생체중아가 될 확률은 13.9% 증가한다.”

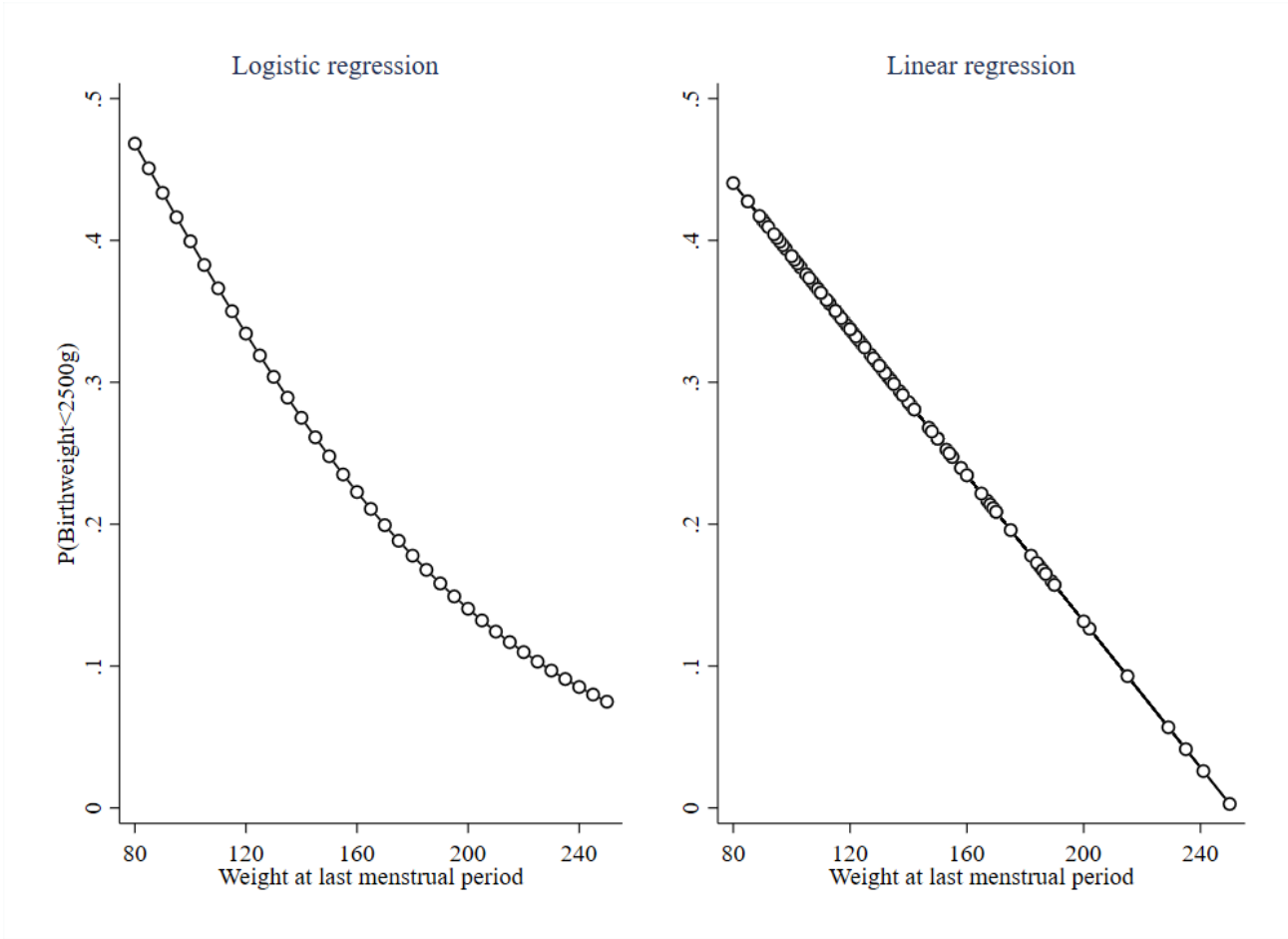
  (3a) “(다른 변수의 영향력을 통제할 때) 산모가 임신중 흡연할 경우 산아가 저출생체중아가 될 오즈는 97.2% 증가한다.”

- 예측확률을 직접 계산해 보면 다음과 같이 해석할 수도 있다(Why?).

  (3b) “(다른 변수의 영향력을 통제할 때) 산모가 임신중 흡연할 경우 산아가 저출생체중아가 될 확률은 25.7%에서 39.3%로 약 13.6% 증가한다.”

유의성 검정

- 선형확률모형과 로지스틱 회귀모형의 추정값은 대체로 비슷하다(늘 그런 것은 아니다). 직접 OLS와 logit의 결과물 차이를 비교해보자.



모형 적합도

모형 적합도

로지스틱 회귀모형의 적합성을 평가하는 것도 중요한 절차가 된다.

- 기존의 실증연구에서는 주로 다음과 같은 적합도 지표를 주로 살펴본다: (1) 우도비 검정(likelihood-ratio test), (2) 유사결정계수(pseudo R^2), (3) 정보기준 (information criterion), 그리고 (4) 분류표(classification table).
- 분류표를 제외한 나머지 지표는 모두 우도함수를 이용하여 계산한다. 로지스틱 회귀모형의 해(solution)는 최대우도법(MLE)으로 계산됨을 기억하자.
- 적합도 지표의 수학적 논리를 이해하기 위해서는 사실 MLE에 대해 좀 더 깊은 이해를 요구한다. MLE는 (사회과학·보건의료 고급통계학 뿐 아니라) 여러 계산사회과학 (computational social science) 분야에서 중요한 기초이다.

모형 적합도

- 첫째, 우도비 검정은 다음의 검정통계량(test statistics) G를 이용한다.

$$G = -2ln\frac{\lambda_{null}}{\lambda_{model}} = -2(ln\lambda_{null} - ln\lambda_{model}) \sim \chi^2_k$$

- λ_{null}은 아무런 독립변수도 넣지 않았을 때 (상수만으로) 얻게 되는 우도함수값이고, λ_{model}은 독립변수를 추가한 경우 얻게 되는 우도함수값이다.
- 우도비 검정의 귀무가설은 다음과 같다.

$$H_0 : \lambda_{null} = \lambda_{model}$$

- 이를 기각하지 못하면 "이 모형은 아무 짝에도 쓸모가 없다"라는 의미로 받아들여진다 (Why?).

모형 적합도

- 똑같은 원리를 사용하여 두 모형을 비교할 때도 우도비 검정을 사용할 수 있다!

$$G = -2ln\frac{\lambda_{restricted}}{\lambda_{full}} = -2(ln\lambda_{restricted} - ln\lambda_{full}) \sim \chi^2_{\Delta k}$$

- $\lambda_{restricted}$는 독립변수가 다소 적게 들어간 모형의 우도함수값이고, λ_{full}은 그보다 독립변수가 좀 더 들어간 모형의 우도함수값이다.
- 이때 제한모형(restricted model)은 완전모형(full model) 안에 내포되어(nested) 있어야 한다.
- χ^2로 검정하는 귀무가설은 아래와 같다.

$$H_0 : \lambda_{restricted} = \lambda_{full}$$

- 이를 기각하지 못하면 "완전모형이 제한모형보다 나은 구석이 없다"라는 의미로 받아들여진다.

모형 적합도

- 둘째, 유사결정계수는 선형회귀모형에서 사용되는 결정계수 R^2를 비선형모형에서 흉내낸 것이다.
- 특히 (1) McFadden's R^2와 (2) Cox-Snell's R^2 가 유명하다.

$$\text{McFadden's } R^2 = 1 - \frac{ln\lambda_{model}}{ln\lambda_{null}}$$

$$\text{Cox-Snell's } R^2 = 1 - \left(\frac{\lambda_{null}}{\lambda_{model}}\right)^{2/n}$$

- 어느 쪽이든 값이 클수록 모형의 설명력이 높다고 말할 수 있다.
- 표본 크기가 충분히 크면 McFadden's R^2는 선형확률모형의 R^2와 제법 유사하다.
- 둘 다 (선형회귀모형의 R^2처럼) 설명된 분산의 비율(proportion of variation explained)로 해석하지 않도록 주의해야 한다(Why?).

모형 적합도

- 셋째, 여러 정보기준 가운데 특히 아카이케 정보기준(Akaike Information Criterion; AIC) 또는 베이즈 정보기준(Bayesian Information Criterion; BIC)이 주로 사용된다.
- AIC는 독립변수의 수 k만 보지만, BIC는 표본 크기 n에도 주목한다.
- AIC는 모형적합도인 $-2LL$에 대해 패러미터(=독립변수+1)의 수 k로 (같은 가중치인 2의) 패널티를 주어 계산한다.

$$AIC = -2ln\lambda_{model} + 2k$$

- 당연히 AIC는 그 값이 작을수록 좋다(Why?).
- 정보기준은 우도비 검정과는 달리 내포성(nestedness) 여부를 따지지 않는다. 그러나 유의성 검정 단계가 없으므로 모형의 개선 여부 판단이 약간 애매모호하다.

모형 적합도

- 넷째, 분류표를 통해 로지스틱 회귀모형에 기반하여 예측된(predicted) $\widehat{Y}$의 {0, 1}와 실제 자료 Y의 {0, 1}를 행렬로 비교하여 나타낸다.
- 이것은 로지스틱 회귀모형의 예측정확성(predictive accuracy)을 평가하는 도구이며 우도함수를 사용하지 않는다.
- 당연히 실제 자료와 모형의 예측이 일치할수록, 즉 정확하게 분류된(correctly classified) 사례가 많을수록 모형이 우수하다고 볼 수 있다.

모형 적합도

예제 3. lowbwt.sav를 이용하여 저출생체중아 여부(low)를 종속변수로, 임신중 흡연 여부(smoke), 산모 연령(age), 임신전 마지막 월경시 산모체중(lwt), 고혈압(ht), 자극성 자궁(ui)을 독립변수로 하는 회귀식을 추정하시오. 그 다음, 인종(race)을 추가한 회귀식을 다시 한 번 추정하고 두 모형을 비교하여 해석하시오(이때 임신중 흡연 여부와 저출생체중 간의 연관성에 초점을 두고 해석하시오).

모형 적합도

- 특히 다중회귀식은 성급하게 추정 단계로 뛰어들어선 안되고 먼저 자료를 꼼꼼히 살펴야 한다.
- 다행히 주어진 자료에는 결측값(missing values)이 없으므로 내포성 문제는 걱정되지 않는다.
- 한편 인종(race)은 범주형 변수이므로 가변수로 바꾸어야 한다([변환]-[더미변수 작성]을 사용한다).
- 추정 결과는 엑셀로 옮겨 정리한다. 구체적인 표의 작성 절차는 좋은 학술지를 그대로 흉내내는 것이 바람직하다.

모형 적합도

- 인종을 고려하지 않았을 때, 흡연은 저출생체중에 대하여 통계적으로 유의한 변수가 아니다(단 $t = 1.92; p = 0.054$). 그러나 인종을 고려하면 큰 차이가 나타난다.
- 두 모형 사이의 우도비 검정 결과 $\chi^2 = 7.83$으로 95% 신뢰수준에서 통계적으로 유의하게 모형의 개선을 확인할 수 있다.
- 유사결정계수도 0.10에서 0.13으로 모형의 개선을 시사한다.
- 마찬가지로 분류표에서는 72.49%에서 74.60%로, AIC에서는 223.8에서 220으로의 개선을 확인할 수 있다.

몇가지 주의사항

몇가지 주의사항

종속변수가 {0, 1}이라고 무조건 로지스틱 회귀분석을 사용할까?

- 사실 이게 그렇게 간단한 문제가 아니다.
- 로지스틱 회귀분석은 완전설정된 모형(fully specified model)에서 설명되고 남은 오차 ϵ이 로지스틱 분포를 따른다는 가정을 전제하고 있다.

$$\epsilon \sim Logistic(0, 1)$$

- 이 가정은 경험적으로 검증될 수 없으며 이로 인해 상당히 복잡한 방법론적 이슈가 발생한다.
- 그러나 이해하지 못해도 괜찮다! 대부분의 리뷰어도 수학을 잘 모르기 때문이다.

몇가지 주의사항

때로는 통계분석 패키지가 로지스틱 회귀식 추정에 실패하기도 한다.

- 최대우도법(MLE)은 β_0와 β_1을 수치해석으로 얻어내기 위해 반복법(iteration method)을 사용하는데, 아무리 반복해도 해를 얻지 못한 경우를 수렴 실패(convergence failures)라고 한다.
- 수렴 실패를 바로잡기 위해서는 일단 모형을 단순화시키고 다시 복잡화의 과정을 밟는 편이 좋다. 이 과정에서 추가 투입하는 변수를 신중하게 선택한다.
- 특히 완전예측(perfect prediction) 또는 분리(separation)가 종종 문제를 일으키므로 독립변수(특히 가변수)를 신중히 살펴보아야 한다(Why?).

몇가지 주의사항

표본 크기가 너무 작으면 분명히 문제가 된다. 하지만!

- 표본의 크기 n에 관해 종종 오해하는데, n 자체도 문제지만 자료에서 성공(1) 사례수가 너무 적으면 이른바 희소사건 편의(rare event bias) 문제를 일으킨다는 지적이 있다.
- 다만 그 대책에 관해서는 약간 논쟁이 있다.
- 또한 생각보다 중요한 이슈 중 하나는 극단값(outliers)의 유무이므로 매우 주의해서 살펴야 한다.

몇가지 주의사항

- 종속변수가 가변수일 때, 로지스틱 회귀분석이 유일한 대안인 것은 아니고, 프로빗(probit) 회귀분석과 보충적 로그로그(complementary log-log; cloglog) 회귀분석같은 대안도 있다.
- 로지스틱 회귀분석과는 종속변수가 살짝 다르다.

$$ln\frac{P(Y=1|X)}{1-P(Y=1|X)} = \beta_0 + \beta_1 X \qquad \text{(logit)}$$

$$\Phi[P(Y=1|X)] = \beta_0 + \beta_1 X \qquad \text{(probit)}$$

$$ln(-ln([1-P(Y=1|X)]) = \beta_0 + \beta_1 X \qquad \text{(cloglog)}$$

몇가지 주의사항

- 그런데 무엇을 사용하건 상당히 비슷한 결과를 얻기 때문에 너무 고민하지 않고 로지스틱을 사용해도 괜찮다.
- 가령 로짓, 프로빗, 보 로그-로그의 누적분포함수(cumulative distribution function; CDF)들을 비교해보면 매우 유사하다.