사회통계연습
종속변수가 명목척도인 경우

김현우, PhD[1]

[1] 충북대학교 사회학과 부교수

진행 순서

1. 질적 종속변수의 분석
2. 종속변수가 가변수인 경우
3. 가변수인 종속변수 분석의 시각화
4. 응용 사례

질적 종속변수의 분석

질적 종속변수의 분석

종속변수가 질적변수라면 회귀분석을 사용할 수 있을까?

- 지난 주 우리는 회귀분석이 가능하긴 한데 보통최소제곱(OLS) 알고리즘으로는 곤란하다고 배웠다.
- 그러나 이렇게 쉽게 포기하기엔 연구나 실무에서 이런 상황이 제법 흔하다.

  (1) "지난 대선 투표에 참여 여부(0=불참, 1=참여)를 설명할 수 있는 개인적 특성과 사회적 조건은 무엇인가?"
  (2) "여가활동의 종류(0=무관심형, 1=관람형, 2=참여형, 3=혼합형)를 예측할 수 있는 사회경제적 속성은 무엇인가?"
  (3) "소득수준은 주관적 행복도(1=매우 불행, 2=다소 불행, 3=불행하지도 행복하지도 않음, 4=다소 행복, 5=매우 행복)에 어떤 영향을 미칠까?"

- 결국 종속변수가 질적변수인 경우도 어떻게든 분석할 수 있어야 한다!

질적 종속변수의 분석

- 종속변수가 질적변수인 경우 다음 세 가지 상황을 상정할 수 있다.

  (1) 종속변수가 가변수인 경우
  (2) 종속변수가 범주형 변수인 경우
  (3) 종속변수가 서열척도인 경우

- 내가 하려는 분석이 어떤 상황인지를 먼저 식별하고 적절한 방식을 선택해야 한다!
- 회귀분석 결과의 시각화 기법을 함께 학습하자!

종속변수가 가변수인 경우

종속변수가 가변수인 경우

종속변수가 가변수인 경우 그냥 OLS 회귀분석을 사용해 볼 수 있다.

- 만일 종속변수 Y가 가변수(0=투표 불참, 1=투표 참여)이고, 독립변수 X가 교육연수라면 다음의 회귀식은 어떤 의미를 가질까?

$$E(Y|X) = \widehat{\beta}_0 + \widehat{\beta}_1 X$$

- "교육연수 X가 한 해 증가할 때 투표 참여 여부 Y는 $\widehat{\beta}_1$ 만큼 증가(또는 감소)한다."
- "교육연수 X가 0일 때, 투표 참여 여부 Y는 $\widehat{\beta}_0$ 이다."
- 이상하게 들릴까? 사실 잘 따져보면 그렇지만도 않다!

종속변수가 가변수인 경우

종속변수가 0 또는 1이라면 그것이 백분율처럼 해석될 수 있기 때문이다!

- 종속변수가 가변수라서 0 또는 1이 아니라, 이게 진짜 양적변수라서 0과 1 사이에 있는 어떤 백분율 값이라고 상상해보자. 그리고 다시 해석해보자.
- "교육연수 X가 한 해 증가할 때 투표 참여 확률 Y는 $\widehat{\beta}_1 \times 100$만큼 증가(또는 감소)한다."
- "교육연수 X가 0일 때, 투표 참여 확률 Y는 $\widehat{\beta}_0 \times 100$이다."
- 단지 참여 여부가 아니라 참여 확률로 바뀌었을 뿐인데, 해석이 제법 자연스러워졌다.

종속변수가 가변수인 경우

- 참여 여부와 참여 확률의 차이는 0과 1 사이에 어떤 값이 존재할 수 있는가에 달려있다.
- 종속변수가 가변수라서 사실 0과 1 사이에 어떤 값을 가질 수 없지만, 마치 그럴수 있는 것처럼 상정하는 접근방법을 선형확률모형(linear probability model)이라고 부른다.

$$E(Y|X) = P(Y = 1|X) = \widehat{\beta}_0 + \widehat{\beta}_1 X$$

- 해석할 때는 아까처럼 참여 여부 대신 참여 확률로 간다!

종속변수가 가변수인 경우

연습 1. lowbwt.sav에서 산모의 마지막 월경시 체중(lwt)을 독립변수로, 영아의 출생시 체중(bwt)을 종속변수로 한 회귀식을 추정하고 그 결과를 해석하시오. 다음으로 같은 독립변수를 사용하되, 저출생체중아 여부(low)를 종속변수로 한 별도의 회귀식을 추정하고 그 결과를 해석하시오. 두 모형을 비교하시오.

종속변수가 가변수인 경우

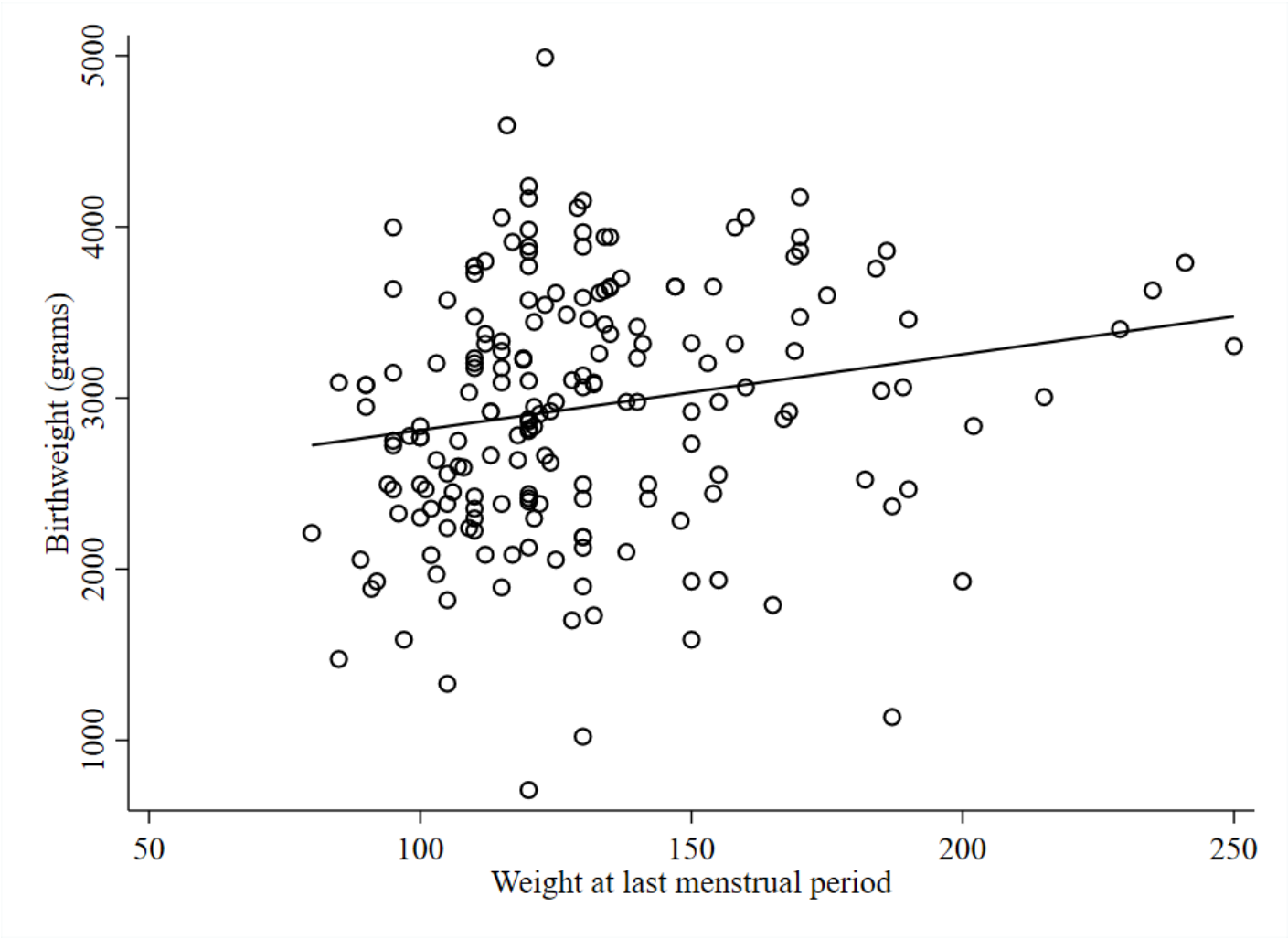
	모형 1 (OLS)		모형 2 (OLS)	
	bwt		low	
엄마의 마지막 생리전 체중	4.430*	(1.713)	-0.003*	(0.001)
상수	2369.2***	(228.5)	0.646***	(0.146)
F	6.686*		5.524*	
결정계수	0.0345		0.0287	
사례수	189			
노트: 괄호 안에 표준오차. * p<0.05, ** p<0.01, *** p<0.001.				

종속변수가 가변수인 경우

- 먼저 bwt를 종속변수로, lwt를 독립변수로 삼아 OLS 회귀분석을 수행한다.
- "산모의 체중이 1 파운드 증가하면 아이의 출생시 체중이 4.43g 증가하고, 이는 95% 신뢰수준에서 통계적으로 유의하다."
- 다음으로 low를 종속변수로, lwt를 독립변수로 삼아 OLS 회귀분석을 수행한다 (선형확률모형).
- "산모의 체중이 1 파운드 증가하면 아이의 출생시 저체중일 확률이 0.3% 감소하고, 이는 95% 신뢰수준에서 통계적으로 유의하다."

종속변수가 가변수인 경우

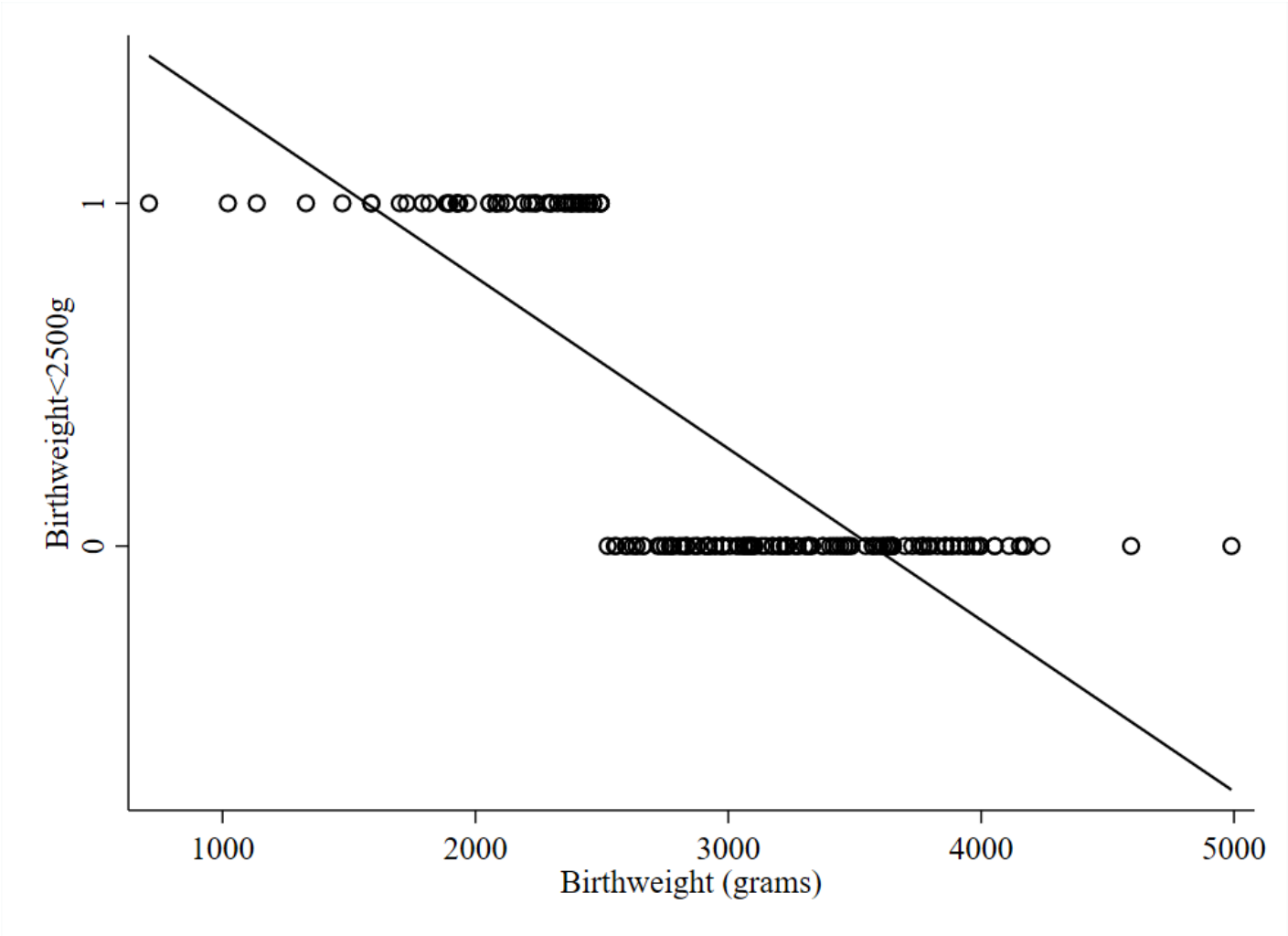
- 우리에게 가장 이상적인 상황은 첫번째처럼 자료가 주어졌을 때이다.



가변수인 종속변수 분석의 시각화

가변수인 종속변수 분석의 시각화

- 하지만 후자로 자료가 주어졌을 수도 있다. 두 변수의 관계는 다음과 같다.



가변수인 종속변수 분석의 시각화

이 상황에서 OLS는 원칙적으로 타당하지 못한 추정법이다.

- 이때는 보통최소제곱(OLS)이 아니라 (우리가 배운 적 없는) 최대우도법(maximum likelihood estimation; MLE)을 통해 회귀계수와 상수를 추정해야 한다.
- 로지스틱 회귀분석(logistic regression analysis)이 바로 최대우도법에 따라 회귀계수와 상수를 추정하는 대표적인 기법이다.
- 함부로 말하기는 조심스럽지만, (특수한 예외를 제외하면) 종속변수가 가변수인 상황에서 OLS 알고리즘과 MLE 알고리즘의 추정량은 결국 비슷한 방향인 경우가 많다.

가변수인 종속변수 분석의 시각화

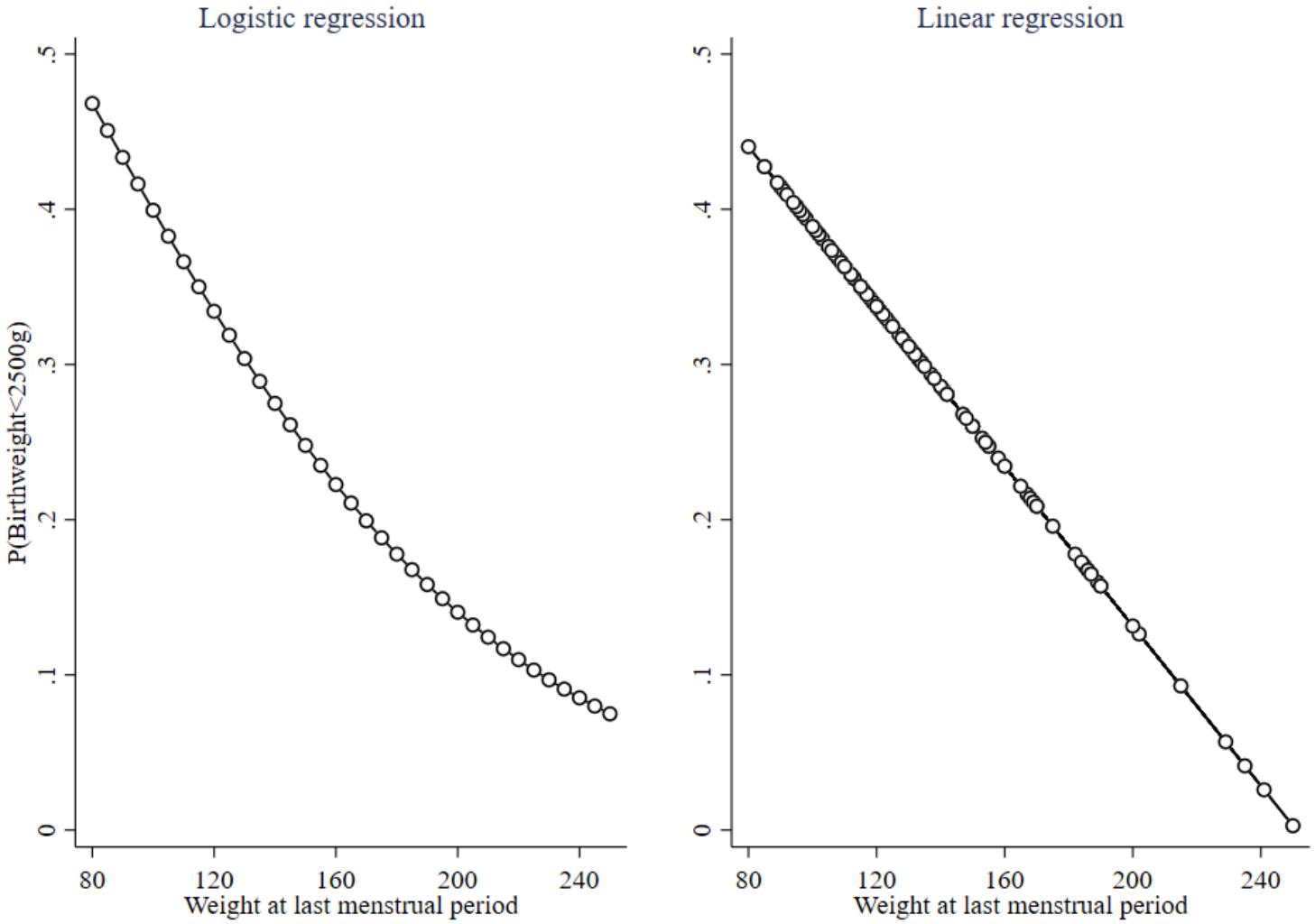
- 선형확률모형을 해석할 때는 단순히 회귀식만 제시하기 보다 (불완전하더라도) 시각화하는 쪽을 추천한다(Why?).
- 시각화를 하려면 먼저 아래처럼 예측확률(predicted probability)을 계산해야 한다.

$$P(Y = 1|X) = \widehat{\beta}_0 + \widehat{\beta}_1 X$$

- 선형회귀분석을 실시하면서 "저장"에서 "비표준화 예측값"을 선택하여 $\widehat{Y}$을 새로 만들어 저장하자!
- 이제 엄마의 마지막 월경시 몸무게(lwt)와 $\widehat{Y}$의 연관성을 보여주는 산점도를 그려보자!

가변수인 종속변수 분석의 시각화

- MLE로 계산된 예측확률과 OLS로 계산된 예측확률 간 상관계수는 0.99이다.



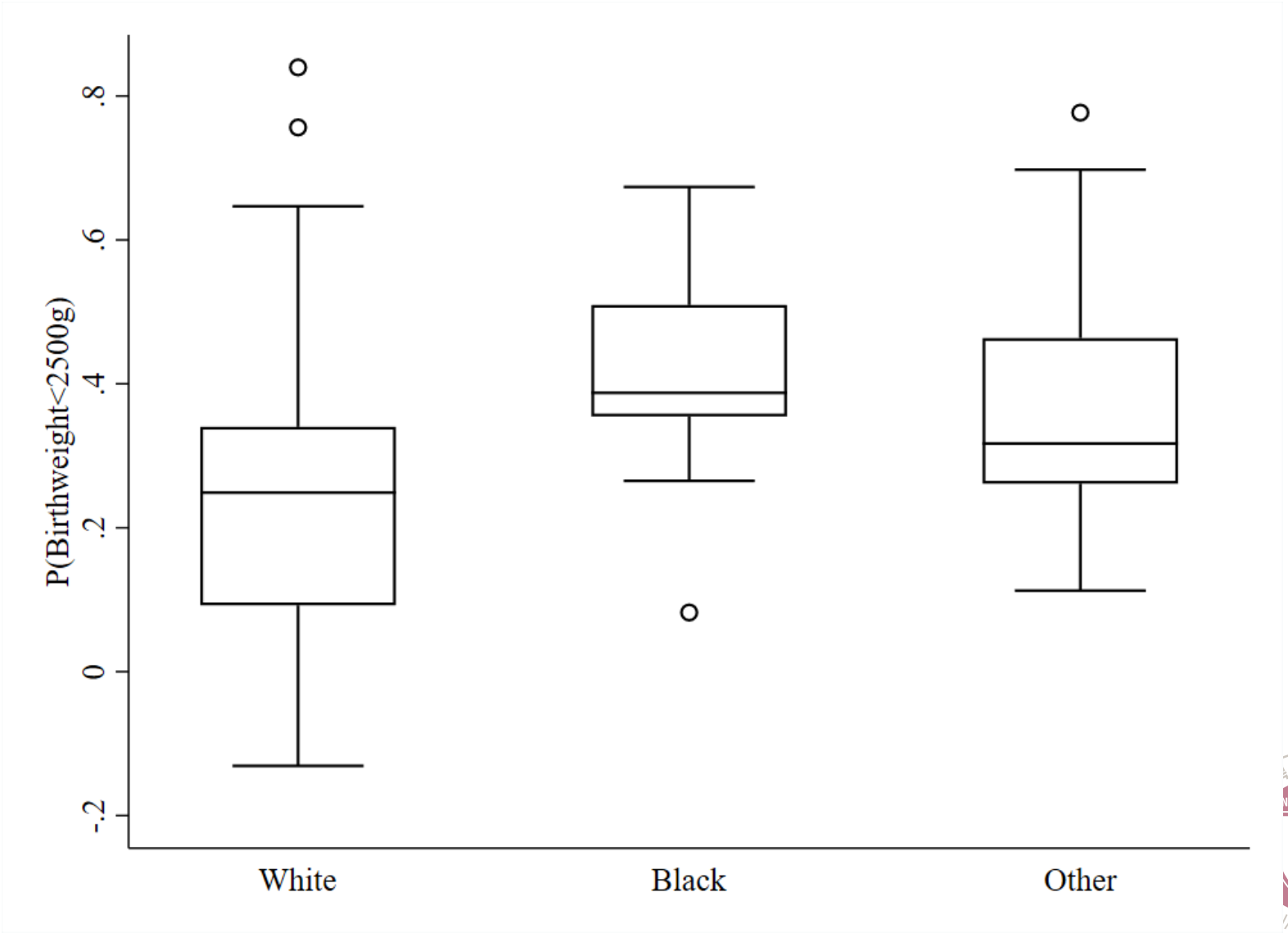
가변수인 종속변수 분석의 시각화

연습 2. lowbwt.sav에서 저출생체중아 여부(low)를 종속변수로 하고, 산모연령(age), 임신중 흡연 여부(smoke), 고혈압(ht), 자극성 자궁(ui), 엄마의 마지막 월경시 체중(lwt), 인종(race), 미숙아 출산력(ptl) 그리고 산과 방문수(ftv)를 독립변수로 하는 회귀식을 추정하고 그 결과를 해석하시오. 인종에 따른 저출생체중아 출산확률을 시각화하시오.

가변수인 종속변수 분석의 시각화

- 우리의 관심변수(variable of interest)인 인종(race)은 범주형 변수이므로 가변수로 바꾸어 회귀분석을 수행해야 한다.
- 그런데 가변수이므로 아까처럼 산점도와 적합선을 그릴 수는 없다(Why?).
- 대신 양적변수와 질적변수의 관계를 살펴보기에 적절한 상자-수염 그림을 그리면 된다!
- 인종(race)과 $\widehat{Y}$ 사이 연관성을 살펴보자.

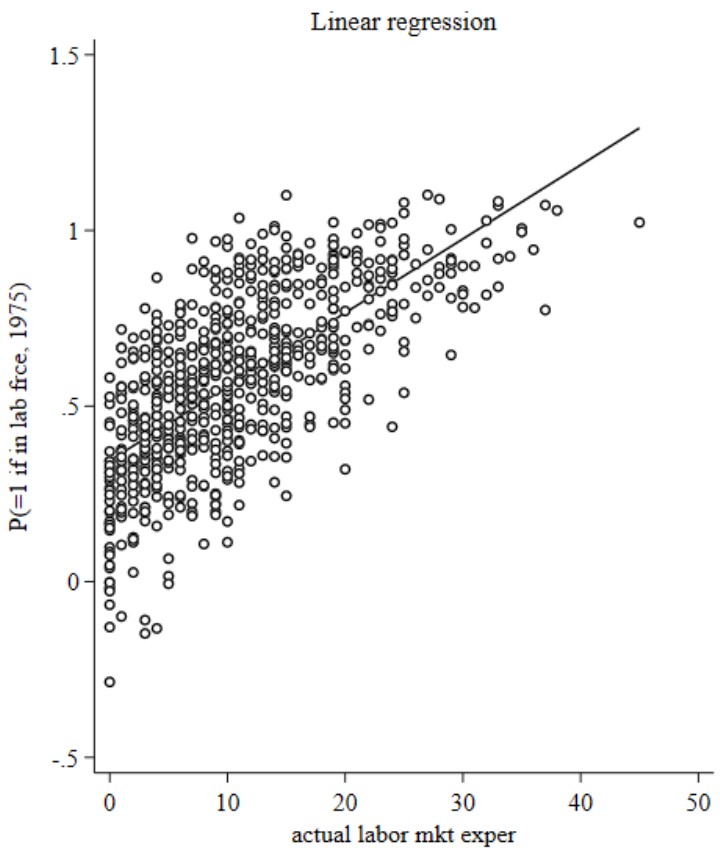
가변수인 종속변수 분석의 시각화

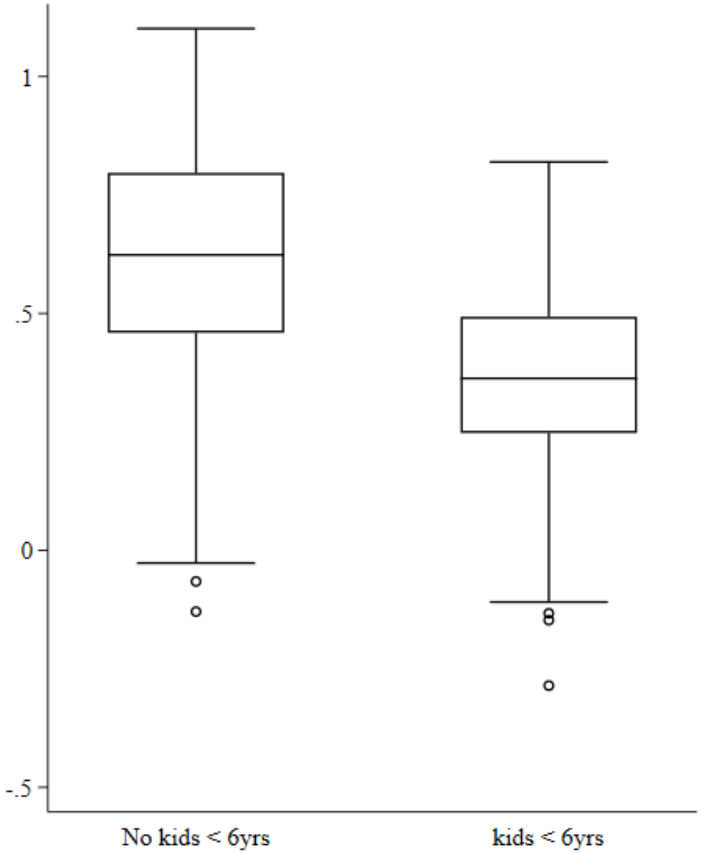


가변수인 종속변수 분석의 시각화

연습 3. mroz.sav에서 아내의 취업 여부(inlf)를 종속변수로 하고, 6세 미만 아이가 있는지 여부(kidslt6), 연령(age), 학력(educ), 남편의 시급(huswage), 도시 거주 여부(city), 일 경험 연수(exper)을 독립변수로 하는 회귀식을 추정하고 그 결과를 해석하시오. 일 경험 연수가 아내의 취업 여부와 어떤 연관성을 가지고 있는지 추정하고 시각화하시오. 6세 미만 아이가 있는지 여부가 아내의 취업 여부와 어떤 연관성을 가지고 있는지 추정하고 시각화하시오.

가변수인 종속변수 분석의 시각화





응용 사례

응용 사례

Dependent Variable: Vote Participation	year=2014				year=2012			
	2-1	2-2	2-3	2-4	2-5	2-6	2-7	2-8
hours worked	-0.001* (0.00)		-0.001* (0.00)		-0.001† (0.00)		-0.001† (0.00)	
overwork		-0.065* (0.03)		-0.060* (0.03)		-0.075** (0.02)		-0.070** (0.03)
Political Psychological Variables and Social Capital Variables								
confidence in the government			0.038 (0.03)	0.040 (0.03)			-0.018 (0.03)	-0.017 (0.03)
political efficiency			0.033** (0.01)	0.033** (0.01)			0.010 (0.01)	0.010 (0.01)
political interest			0.117*** (0.02)	0.115*** (0.02)			0.039*** (0.01)	0.040*** (0.01)
political inclination			-0.020† (0.01)	-0.019† (0.01)			0.011 (0.01)	0.012 (0.01)
political satisfaction			0.013 (0.01)	0.012 (0.01)			-0.010 (0.01)	-0.010 (0.01)
social trust			0.010 (0.01)	0.011 (0.01)			0.008 (0.02)	0.008 (0.02)
social network			-0.002 (0.01)	-0.001 (0.01)			0.000 (0.01)	0.000 (0.01)
social participation			0.105† (0.06)	0.107† (0.06)			0.297*** (0.08)	0.293*** (0.08)
Demographic Variables								
female (ref : male)	0.009 (0.03)	0.010 (0.03)	-0.012 (0.03)	-0.013 (0.03)	0.026 (0.03)	0.028 (0.03)	-0.008 (0.03)	-0.006 (0.03)
educational level(ref: below middle school)								
high school	-0.088* (0.04)	-0.089* (0.04)	-0.102* (0.04)	-0.102** (0.04)	-0.043 (0.03)	-0.042 (0.03)	-0.089* (0.04)	-0.088* (0.04)
above college	0.043 (0.04)	0.038 (0.04)	-0.020 (0.04)	-0.024 (0.04)	0.094* (0.04)	0.088* (0.04)	0.030 (0.04)	0.025 (0.04)
age group(ref: 20s)								
age group: 30s	-0.021 (0.06)	-0.022 (0.06)	-0.012 (0.06)	-0.016 (0.06)	0.116† (0.06)	0.116† (0.06)	0.125† (0.06)	0.125† (0.06)
age group: 40s	0.109† (0.06)	0.112† (0.06)	0.094 (0.06)	0.093 (0.06)	0.143* (0.07)	0.142* (0.07)	0.105 (0.07)	0.103 (0.07)
age group: 50s	0.198** (0.07)	0.200** (0.07)	0.147* (0.07)	0.146* (0.07)	0.211** (0.08)	0.210** (0.08)	0.156† (0.08)	0.153† (0.08)
age group: 60s	0.184* (0.07)	0.187** (0.07)	0.128† (0.07)	0.127† (0.07)	0.314*** (0.08)	0.311*** (0.08)	0.266** (0.08)	0.262** (0.08)

(o be continued))

응용 사례

marital status(ref: married)								
married but no spouse	-0.084† (0.05)	-0.088† (0.05)	-0.088† (0.05)	-0.091* (0.05)	-0.071† (0.04)	-0.072† (0.04)	-0.035 (0.04)	-0.035 (0.04)
non married	-0.114† (0.06)	-0.110† (0.06)	-0.118* (0.06)	-0.116† (0.06)	-0.074 (0.07)	-0.071 (0.07)	-0.094 (0.07)	-0.093 (0.07)
work status dummy & job dummy								
regular	0.011 (0.03)	0.003 (0.03)	0.027 (0.03)	0.015 (0.03)	0.028 (0.04)	0.021 (0.03)	0.028 (0.04)	0.020 (0.04)
irregular	-0.096† (0.05)	-0.107* (0.05)	-0.104* (0.05)	-0.115* (0.05)	-0.074 (0.05)	-0.081† (0.05)	-0.032 (0.05)	-0.040 (0.05)
employer	-0.003 (0.04)	-0.008 (0.04)	0.013 (0.04)	0.005 (0.04)	0.002 (0.04)	-0.002 (0.04)	0.011 (0.04)	0.007 (0.04)
household characteristics								
double-income family	-0.017 (0.03)	-0.019 (0.03)	-0.036 (0.03)	-0.038 (0.03)	0.031 (0.04)	0.031 (0.04)	0.018 (0.04)	0.017 (0.04)
preschool child	-0.026 (0.04)	-0.026 (0.04)	-0.012 (0.04)	-0.014 (0.05)	-0.051 (0.05)	-0.051 (0.05)	-0.071 (0.06)	-0.072 (0.06)
house income	0.007 (0.01)	0.008 (0.01)	0.004 (0.01)	0.004 (0.01)	0.001 (0.01)	0.001 (0.01)	0.003 (0.01)	0.003 (0.01)
urban	-0.019 (0.04)	-0.020 (0.04)	-0.046 (0.04)	-0.047 (0.04)	0.048 (0.03)	0.049 (0.03)	0.007 (0.03)	0.008 (0.03)
region: capital	-0.042 (0.03)	-0.038 (0.03)	-0.072* (0.03)	-0.067* (0.03)	-0.009 (0.04)	-0.008 (0.04)	-0.007 (0.04)	-0.006 (0.04)
region: Honam	-0.036 (0.04)	-0.034 (0.04)	-0.062 (0.04)	-0.059 (0.04)	0.036 (0.04)	0.035 (0.04)	-0.004 (0.05)	-0.004 (0.05)
region: Youngnam	-0.090* (0.04)	-0.087* (0.04)	-0.125*** (0.04)	-0.123*** (0.04)	-0.000 (0.04)	0.003 (0.04)	-0.039 (0.04)	-0.036 (0.04)
Constant	0.738*** (0.09)	0.721*** (0.09)	0.464*** (0.11)	0.443*** (0.11)	0.505*** (0.10)	0.503*** (0.10)	0.411** (0.14)	0.407** (0.14)
Observations	1,337	1,337	1,313	1,313	1,377	1,377	1,188	1,188
R-squared	0.080	0.082	0.136	0.136	0.071	0.074	0.111	0.114
F	5.263	5.414	8.550	7.863	4.729	4.924	5.402	5.617
p	.000	.000	.000	.000	.000	.000	.000	.000

Robust standard errors in parentheses. *** $p<0.001$, ** $p<0.01$, * $p<0.05$, † $p<0.1$

신영민. 2019. "정치적 참여의 자원으로서 노동시간에 관한 탐색적 연구." 『한국사회학』 53(2): 1-42.