10. 다중회귀분석 II

정치와 데이터분석

박상훈 (sh.park.poli@gmail.com)
강원대학교

오늘의 목표

.pull-left[

10:05-10:45

다중회귀분석의 기초 복습

11:00-11:40

다중회귀분석에서 상호작용항의 이해

11:55-12:35

실습과제 해설 및 질의응답

Part I. 다중선형회귀모형의 복습

다중회귀분석 II

OLS 추정의 Gauss-Markov 가정 (MLR 확장)

MLR 1. 선형성 (Linearity): $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + u$

MLR 2. 무작위 표본 (Random Sample): 데이터가 모집단으로부터 무작위로 추출됨.

MLR 3. 내생성 없음 (No Endogeneity)/외생성 (Exogeneity): 오차항 u는 모든 설명변수 $X_1, \ldots, X_k$와 독립적. $E(u|X_1, \ldots, X_k) = 0$. 즉, OVB가 없어야 함을 의미

MLR 4. 오차항의 동분산성 (Homoskedasticity): $Var(u|X_1, \ldots, X_k) = \sigma^2$ (일정)

MLR 5. 완벽한 다중공선성 없음 (No Perfect Multicollinearity)

- 설명변수들 간에 *완벽한* 선형 관계가 존재해서는 안 됨.
- 완벽한 다중공선성 (Perfect Multicollinearity): 설명변수 중 하나가 다른 설명변수(들)의 조합으로 완벽하게 표현되는 경우.

다중회귀분석 II

OLS 추정의 Gauss-Markov 가정 (MLR 확장)

MLR 1. 선형성 (Linearity)

MLR 2. 무작위 표본 (Random Sample)

MLR 3. 내생성 없음 (No Endogeneity)/외생성 (Exogeneity)

MLR 4. 오차항의 동분산성 (Homoskedasticity)

MLR 5. 완벽한 다중공선성 없음 (No Perfect Multicollinearity)

⇒ **위 가정들이 충족되면 OLS 추정치는 편향되지 않고(unbiased), 효율적(최소분산)이며, Gaus-Markov 정리에 따라 BLUE**

다중회귀분석 II

다중회귀계수의 해석

다중회귀모형: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i$

각 계수 β_j는 다른 모든 변수들을 일정하게 유지한 상태에서 X_j가 1 단위 증가할 때 Y의 변화량 (부분효과)

β_1 = X_1의 한 단위 증가가 기대 Y에 미치는 영향 (다른 X가 동일할 때)

다중회귀에서는 상관관계를 통제한 효과를 추정: 한 변수의 계수는 그 변수의 순수한 영향 (다른 변수들의 영향 제거 후)

다중회귀분석 II

범주형 변수의 처리: 더미변수와 기준 범주

범주형 독립변수는 더미변수(0/1)들로 변환하여 회귀에 포함

- 범주 개수가 m이면 $m-1$개의 더미변수를 사용하고 나머지 하나를 기준 범주로 설정
- 기준 범주의 효과는 절편에 흡수되며, 각 더미변수의 계수는 해당 범주와 기준 범주 간의 Y 차이를 나타냄(다른 변수 통제하에).
    - 성별 변수(남/여) $\rightarrow$ Female = 1 더미 (기준: Male = 0).
    - 회귀식 $Y = \beta_0 + \beta_1 X + \beta_2 \text{Female} + u$
        - β_2: Female과 Male의 평균적 Y 차이 (다른 변수 통제 시)
        - β_0: 기준범주(Male)의 Y 절편 (Female = 0인 경우의 예상 Y)

다중회귀분석 II

누락변수 편의 (Omitted Variable Bias)

회귀모형에 포함되어야 할 중요한 설명변수를 빠뜨리면, 추정된 계수들이 편향될 수 있음.

누락변수 Z가 Y에 영향을 주고 포함된 X와 상관되어 있으면, X의 계수에 Z의 효과가 섞여 잘못 추정됨(상관된 부분을 X가 설명한 것으로 착각).

- OVB 조건: (1) 누락변수가 Y에 유의미한 영향, (2) 누락변수가 포함된 독립변수와 상관 $\rightarrow$ 이 경우 OLS 추정치는 일관된(consistently) 추정이 어려움.
- 결과: 계수의 부호/크기가 실제와 다르게 추정될 위험. 편의를 줄이려면 관련 변수를 모두 모형에 포함시키는 것이 중요
    - 교육수준을 뺀 채 근로경력(X)이 임금(Y)에 미치는 효과를 추정하면, 교육과 경력이 상관되므로 경력 계수에 교육의 효과가 섞여 편향 발생.

다중회귀분석 II

결정계수 R^2와 수정 R^2

결정계수 R^2: 종속변수 Y 변동 중 회귀모형이 설명하는 비율. $0 \leq R^2 \leq 1$

- $R^2 = 1 - \frac{\text{SSR}}{\text{SST}}$ (SSR: 잔차제곱합, SST: 총변동)
- R^2 높을수록 모형이 데이터를 잘 설명 (적합도 높음)

수정된 결정계수 $\bar{R}^2$: 설명변수 개수를 고려하여 R^2를 보정한 지표

- $\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1}$ (n: 표본수, p: 추정 파라미터 수), R^2_{adj} 라고도 함.
- 불필요한 변수를 추가하면 $\bar{R}^2$는 감소할 수 있음 (과적합 페널티)
- 변수 개수가 다른 모형들 비교 시 사용. 가장 높은 $\bar{R}^2$ 갖는 모형이 균형적 설명력 좋음

다중회귀분석 II

F-검정: 모형 전체 유의성

F-통계량: 회귀모형에서 전체 계수의 유의성을 검정하는 데 사용

- 귀무가설 H_0: 모든 설명변수의 계수가 0 ($\beta_1 = \beta_2 = \cdots = \beta_k = 0$)
- 대립가설 H_1: 적어도 하나의 계수가 0이 아니다 (모형에 설명력 있음)
- F = $\frac{(SSR_{H_0} - SSR_{\text{full}})/k}{SSR_{\text{full}}/(n-k-1)}$ ~ $F_{k,,n-k-1}$

$p-$값이 작으면 (일반적으로 0.05 이하) H_0 기각 $\rightarrow$ 모형 전체적으로 유의미하다고 결론

- 적어도 하나의 독립변수는 종속변수에 통계적으로 유의한 영향 있음(모형이 빈 모형 대비 유의한 설명력 가짐).
- F-검정은 특정 변수들의 집합에 대한 검정도 가능(더미변수 여러 개의 공동 유의성 검정)

다중회귀분석 II

다중공선성과 VIF

다중공선성

- 둘 이상의 독립변수들이 강한 선형관계를 가져 상대적 변별이 어렵게 되는 현상
- 심한 경우 계수 추정 불안정, 표준오차 급증 $\rightarrow$ 통계적 유의미성 감소

완전 공선성: 독립변수들이 정확한 선형결합 관계에 있을 때 (계수 추정 불가능)

- 다중공선성은 추정치의 편의는 유발하지 않지만, 표준 오차를 키워 신뢰구간을 넓히고 회귀 계수의 통계적 유의확률을 낮춤
- 따라서 중요한 변수가 유의하지 않게 나타날 수 있음; 데이터 약간만 변화해도 계수값이 크게 달라질 수 있음

다중회귀분석 II

다중공선성과 VIF

VIF(Variance Inflation Factor): 다중공선성 정도를 수치화한 지표

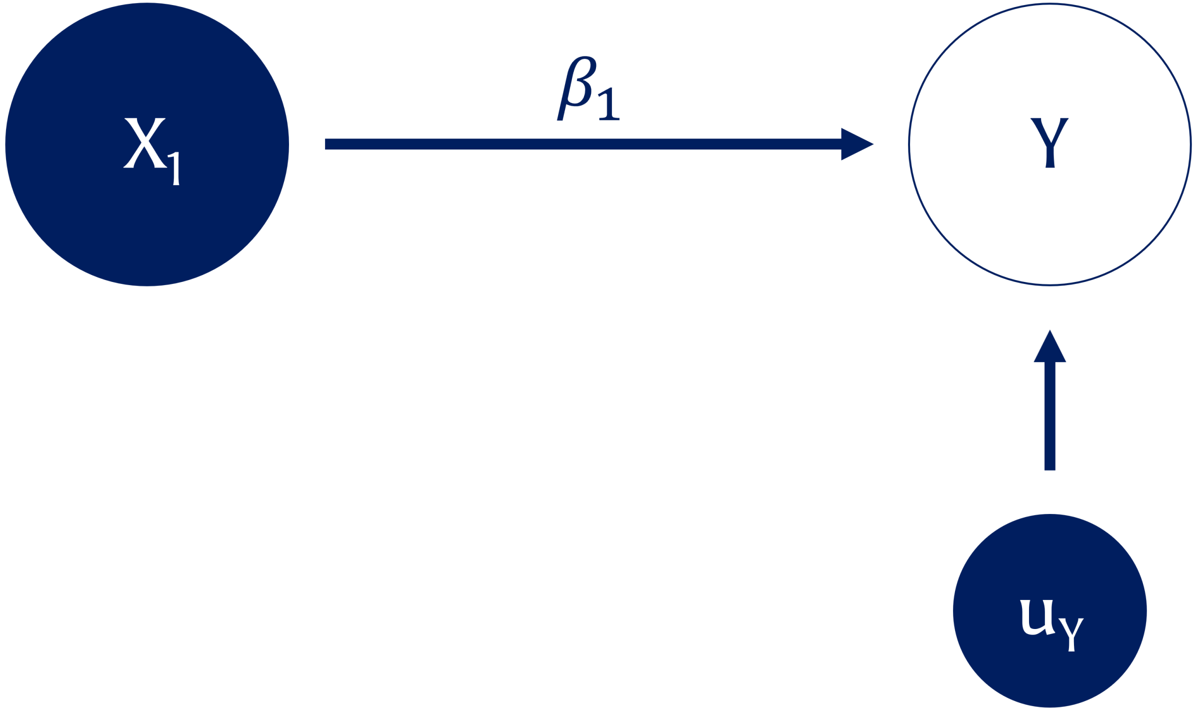
- $VIF_j = \frac{1}{1-R_j^2}$ (R_j^2: X_j를 다른 독립변수들로 회귀한 결정계수)
- VIF_j가 높을수록 X_j가 다른 변수들과 강한 상관 $\rightarrow$ 계수분산 증가
- 일반적 기준: $VIF > 10$이면 심각한 다중공선성 의심 (혹은 $ >5$도 주의)

높은 상관의 변수 제거 또는 주성분/축소기법 활용, 표본 확대 등으로 관리

다중회귀분석 II

기본 모형

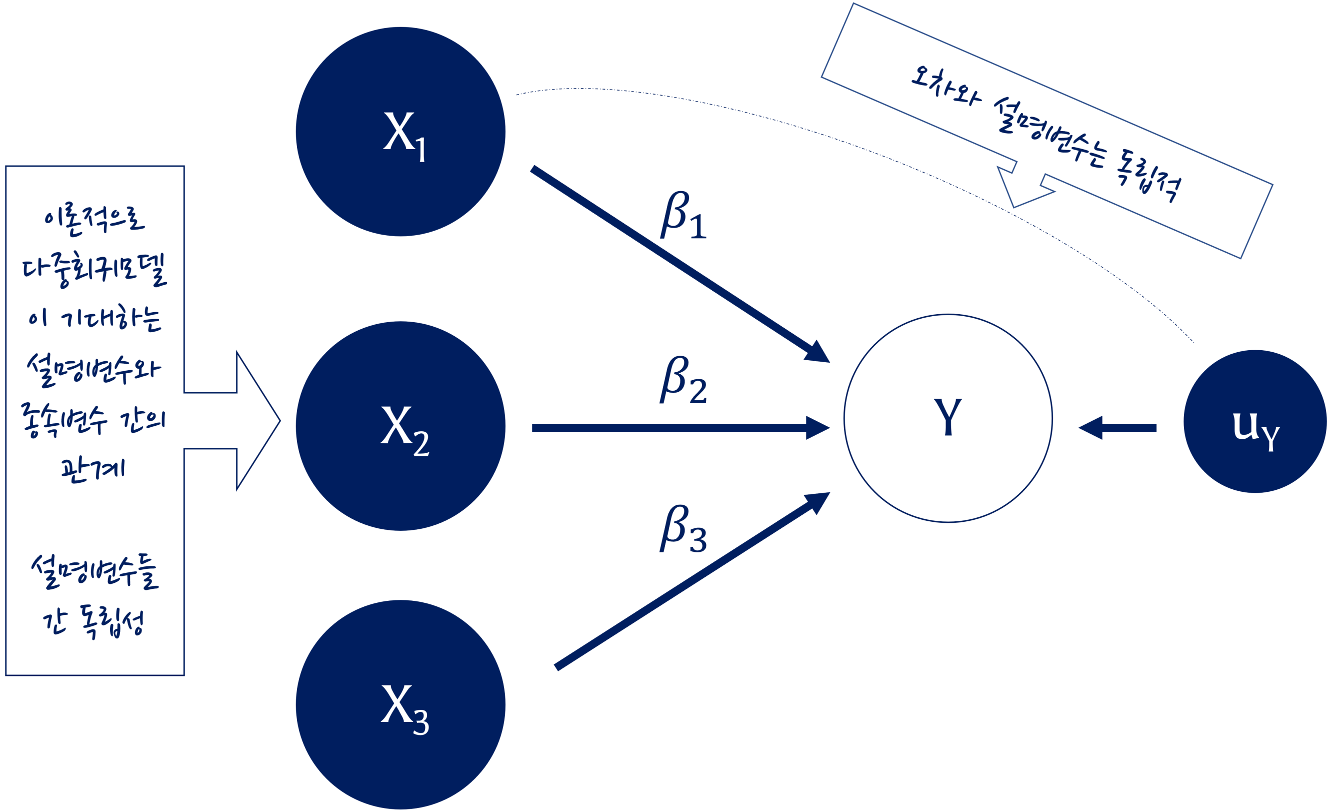
$$\boxed{Y = \beta_0 + \beta_1 X_1 + u}$$



다중회귀분석 II

기본 모형

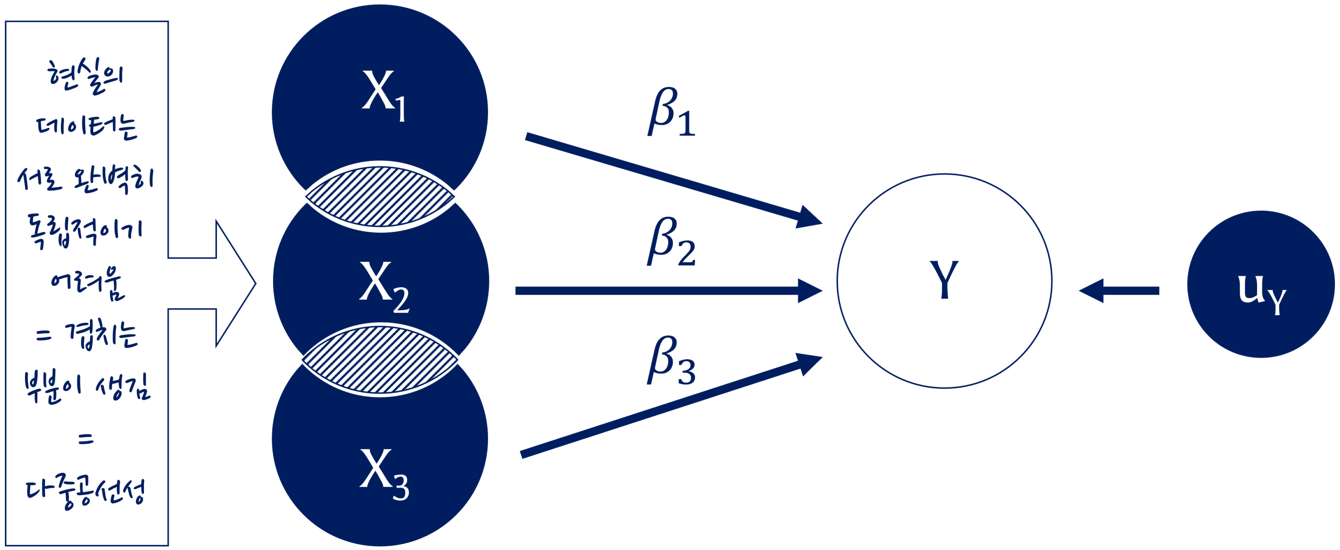
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + u$$



다중회귀분석 II

기본 모형

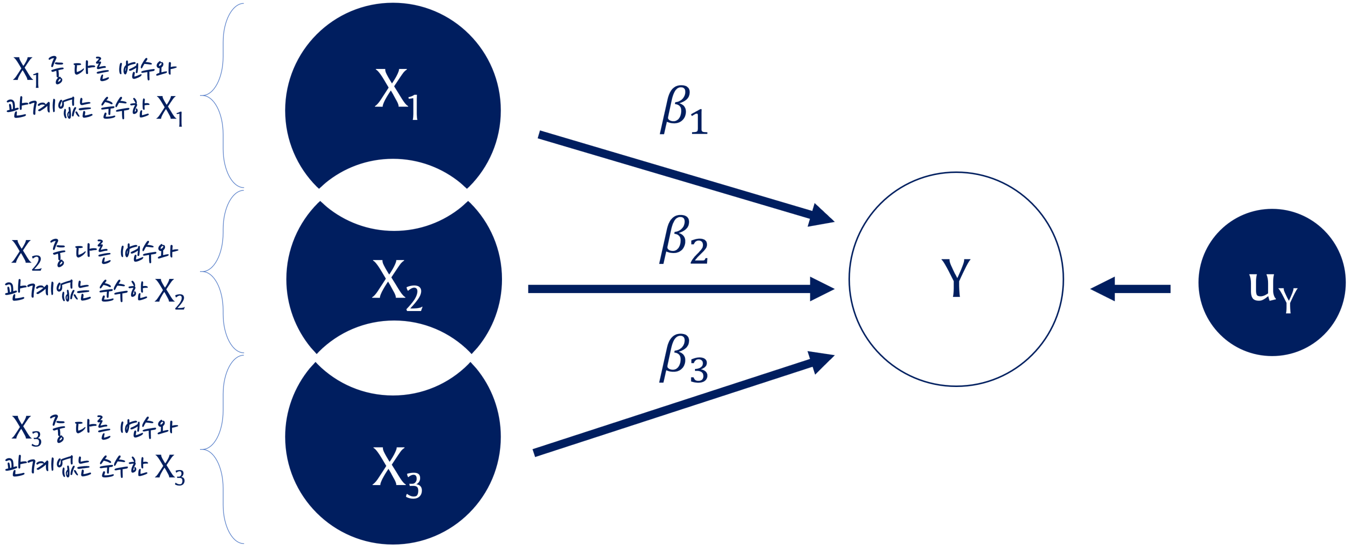
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + u$$



다중회귀분석 II

기본 모형

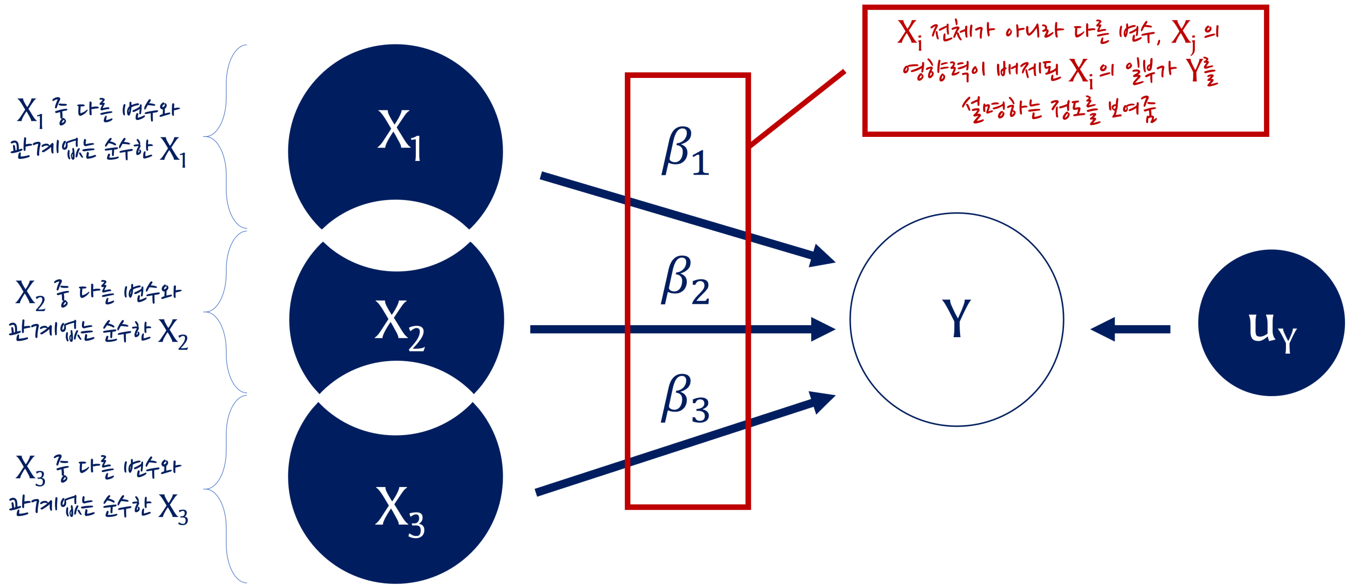
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + u$$



다중회귀분석 II

기본 모형

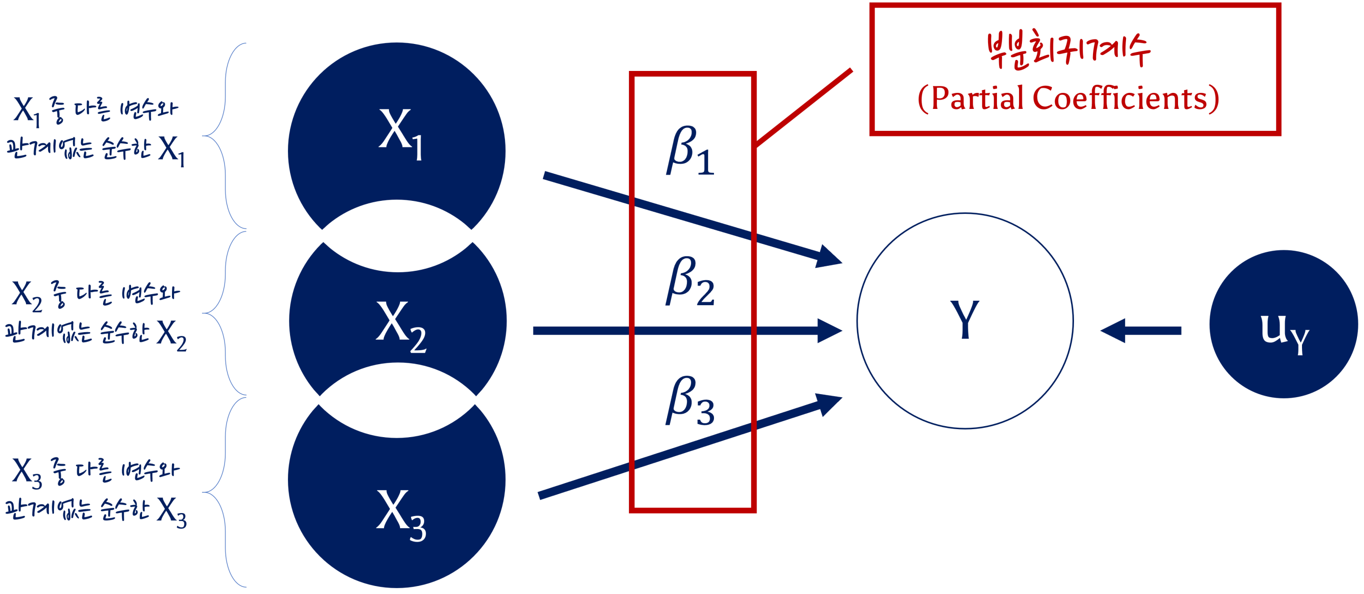
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + u$$



다중회귀분석 II

기본 모형

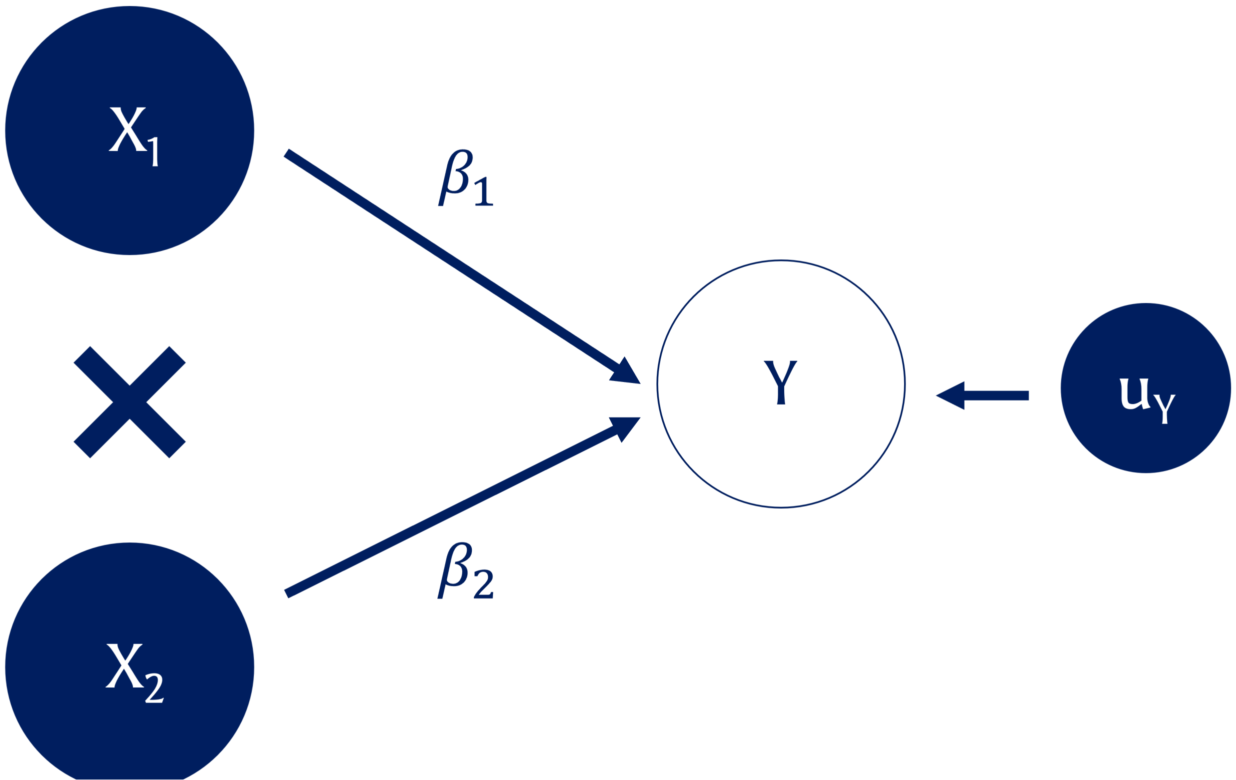
$$\boxed{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + u}$$



다중회귀분석 II

상호작용효과: 기본모형

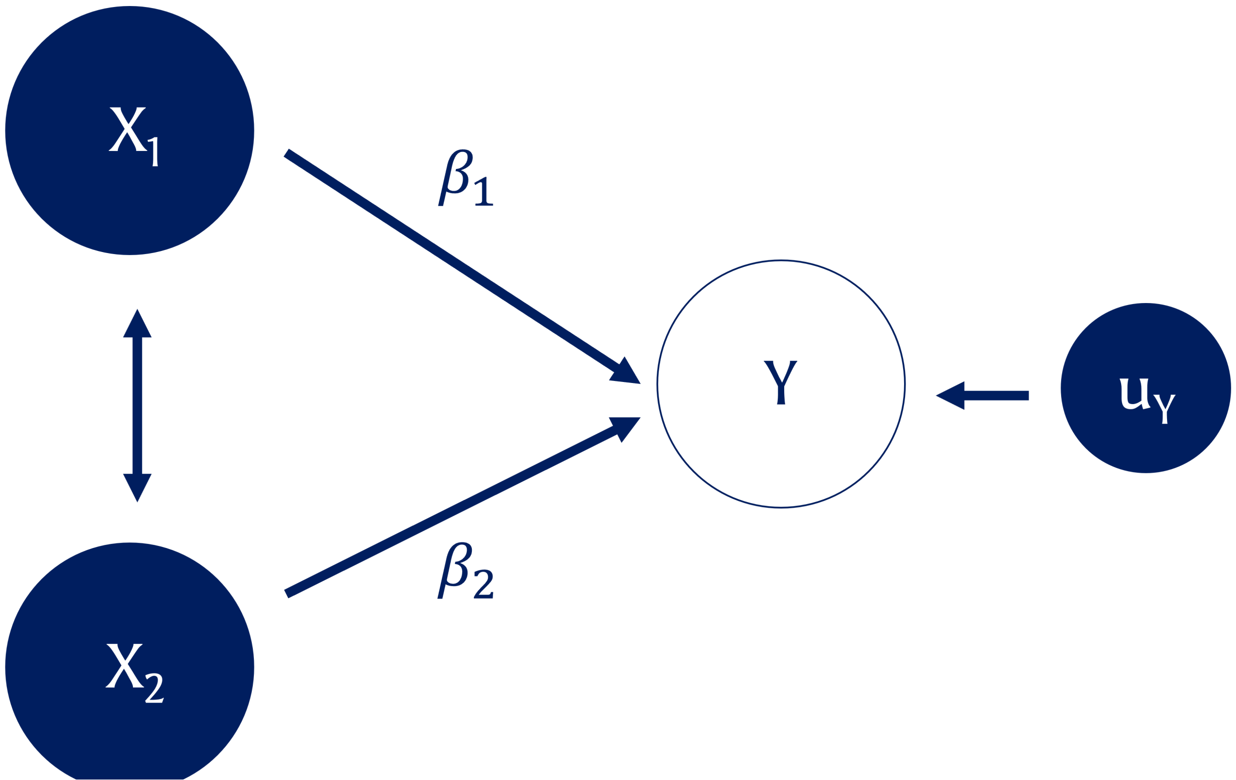
$$\boxed{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + u}$$



다중회귀분석 II

상호작용효과: 기본모형

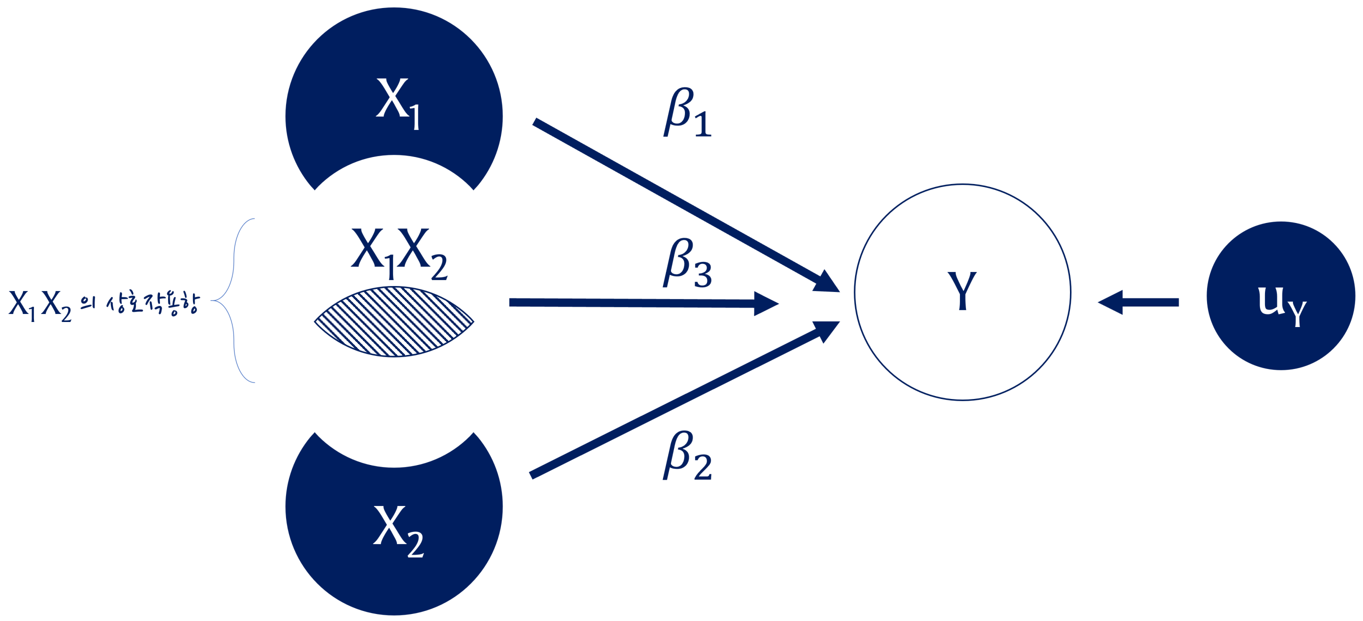
$$\boxed{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + u}$$



다중회귀분석 II

상호작용효과: 기본모형

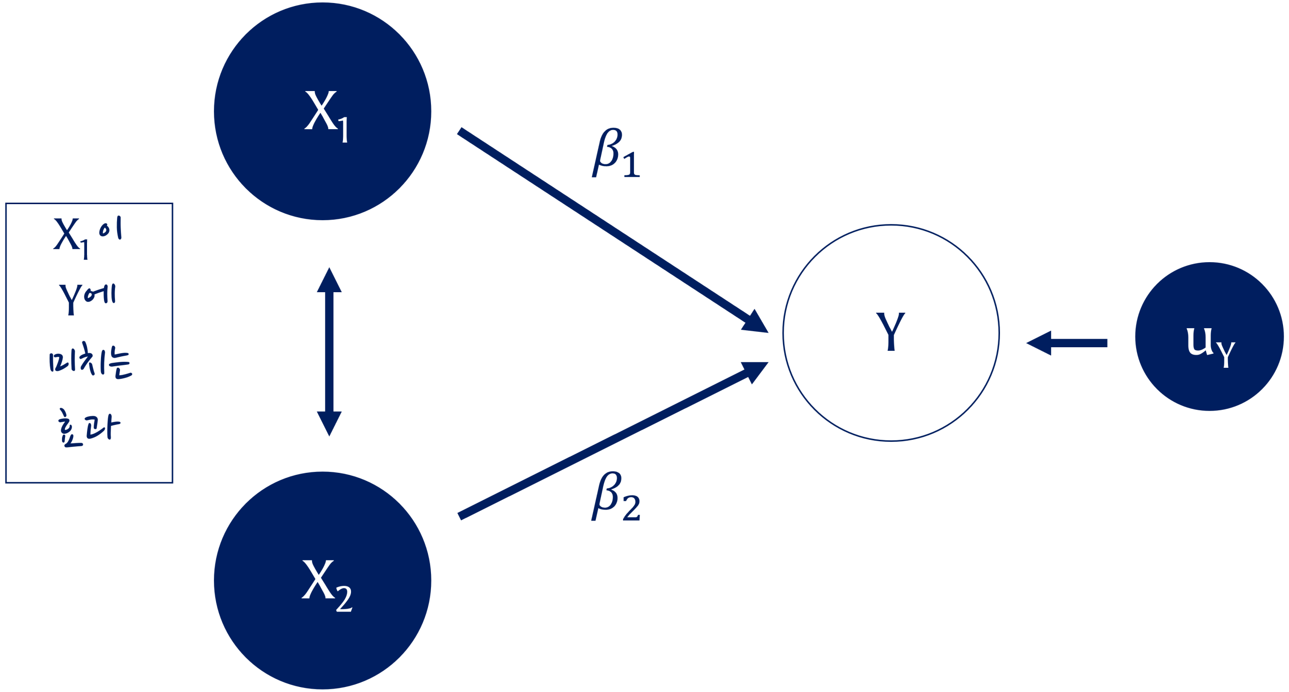
$$\boxed{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + u}$$



다중회귀분석 II

상호작용효과: 기본모형

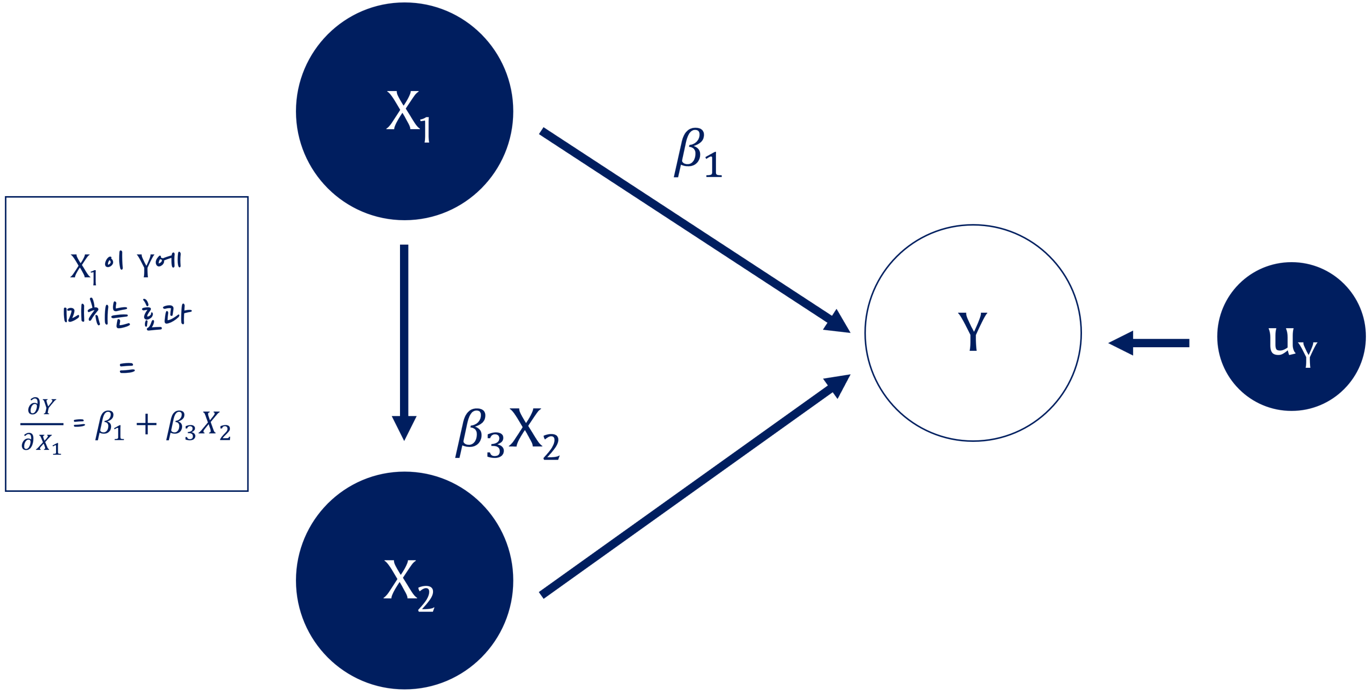
$$\boxed{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + u}$$



다중회귀분석 II

상호작용효과: 기본모형

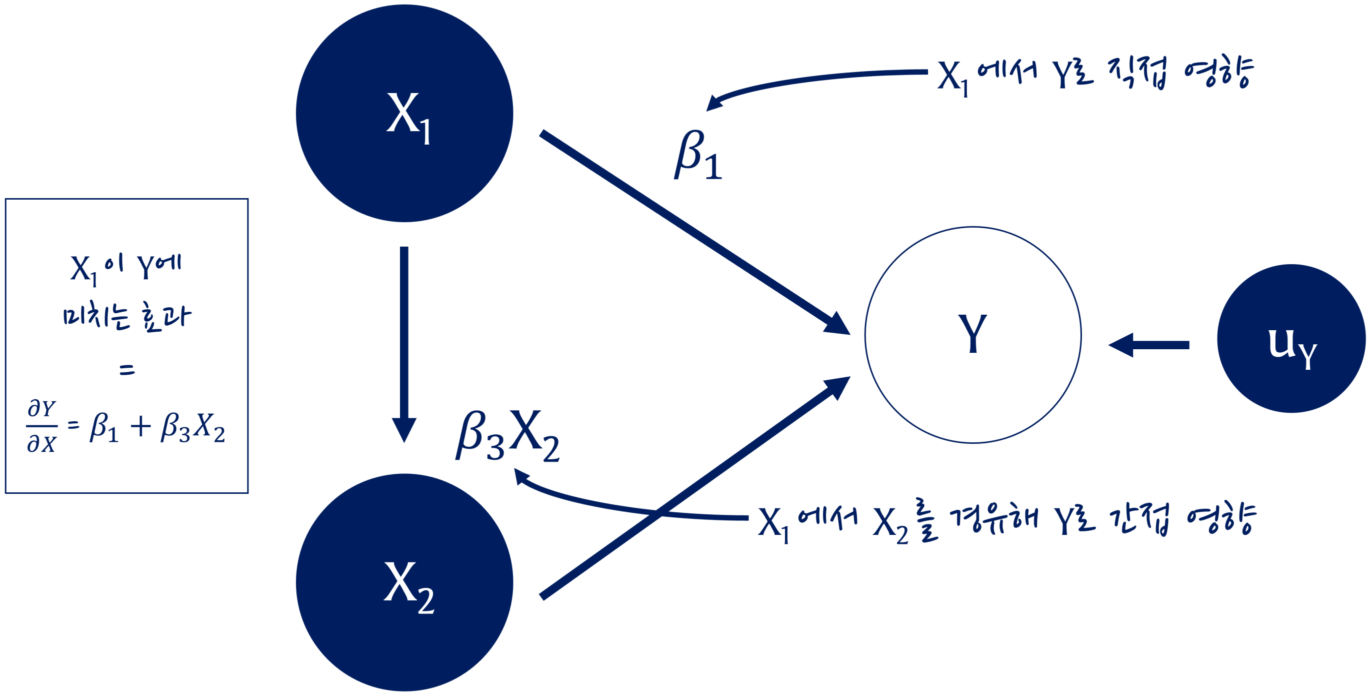
$$\boxed{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + u}$$



다중회귀분석 II

상호작용효과: 기본모형

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + u$$

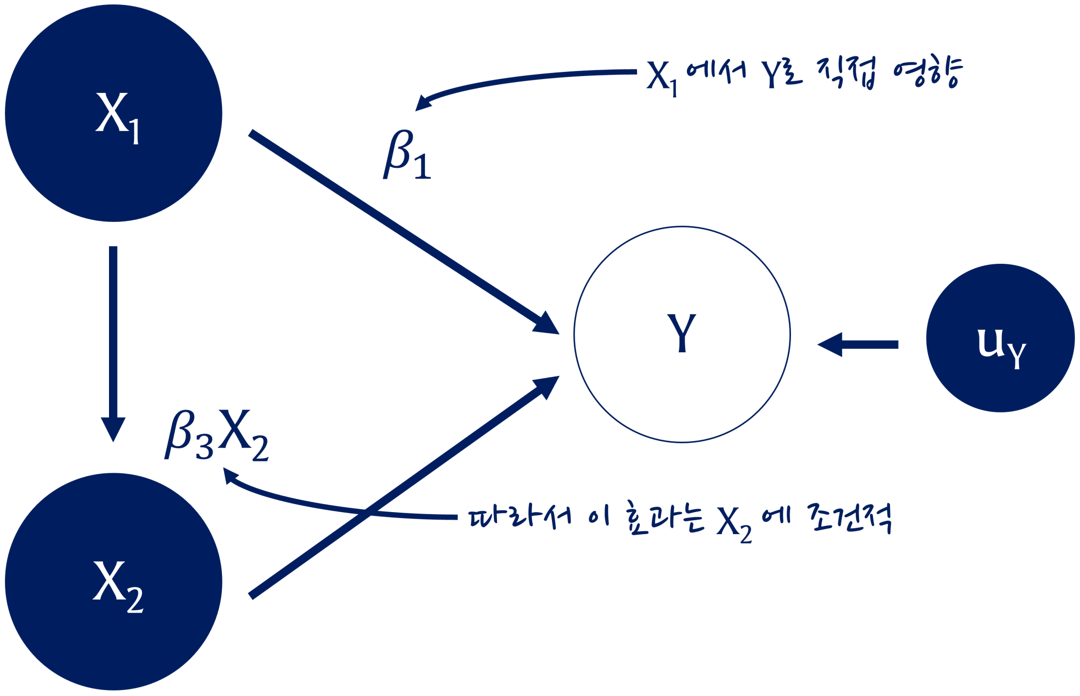


다중회귀분석 II

상호작용효과: 기본모형

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + u$$

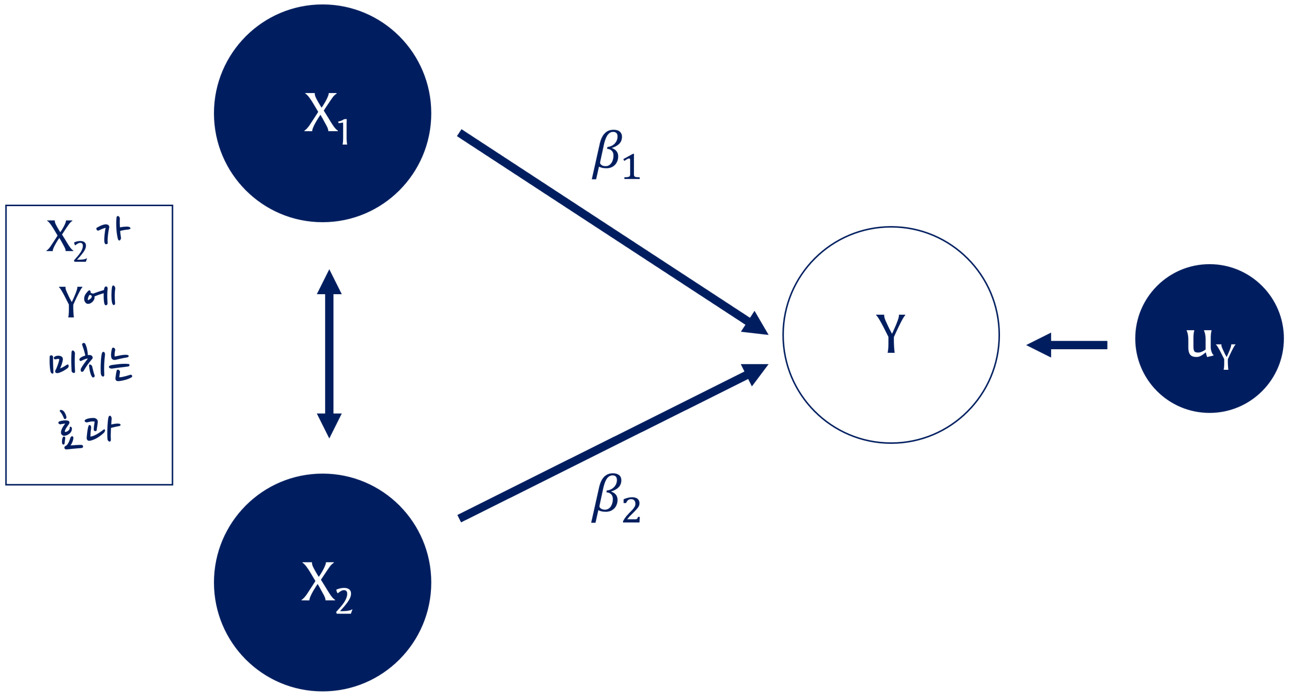
X_1이 Y에
미치는 효과
=
$\frac{\partial Y}{\partial X} = \beta_1 + \beta_3 X_2$



다중회귀분석 II

상호작용효과: 기본모형

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + u$$

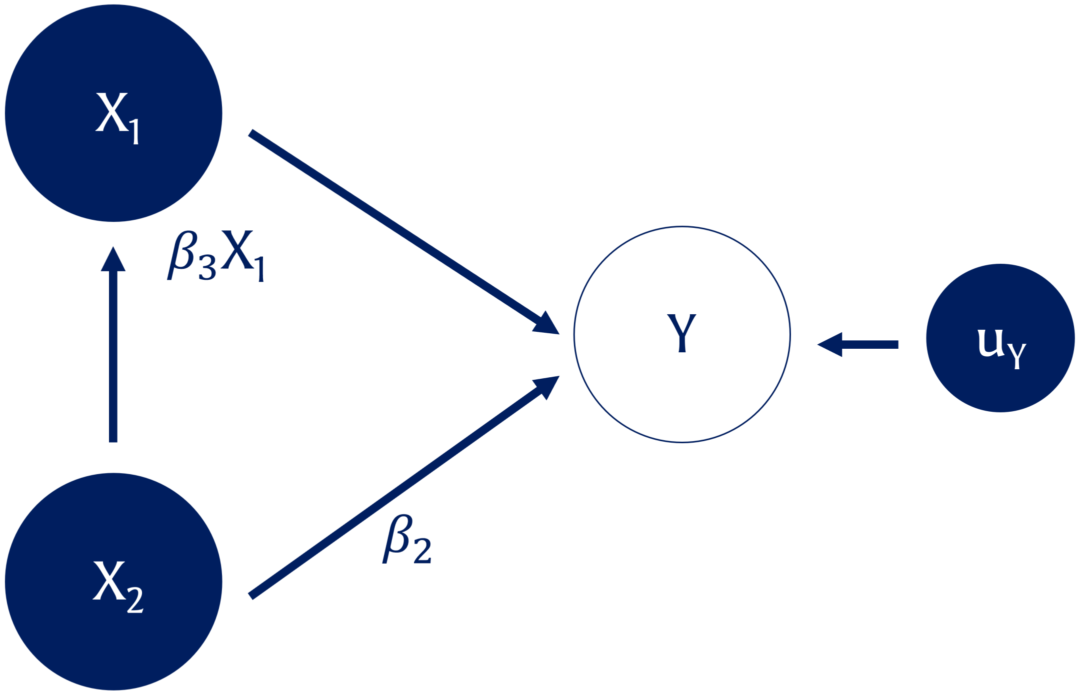


다중회귀분석 II

상호작용효과: 기본모형

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + u$$

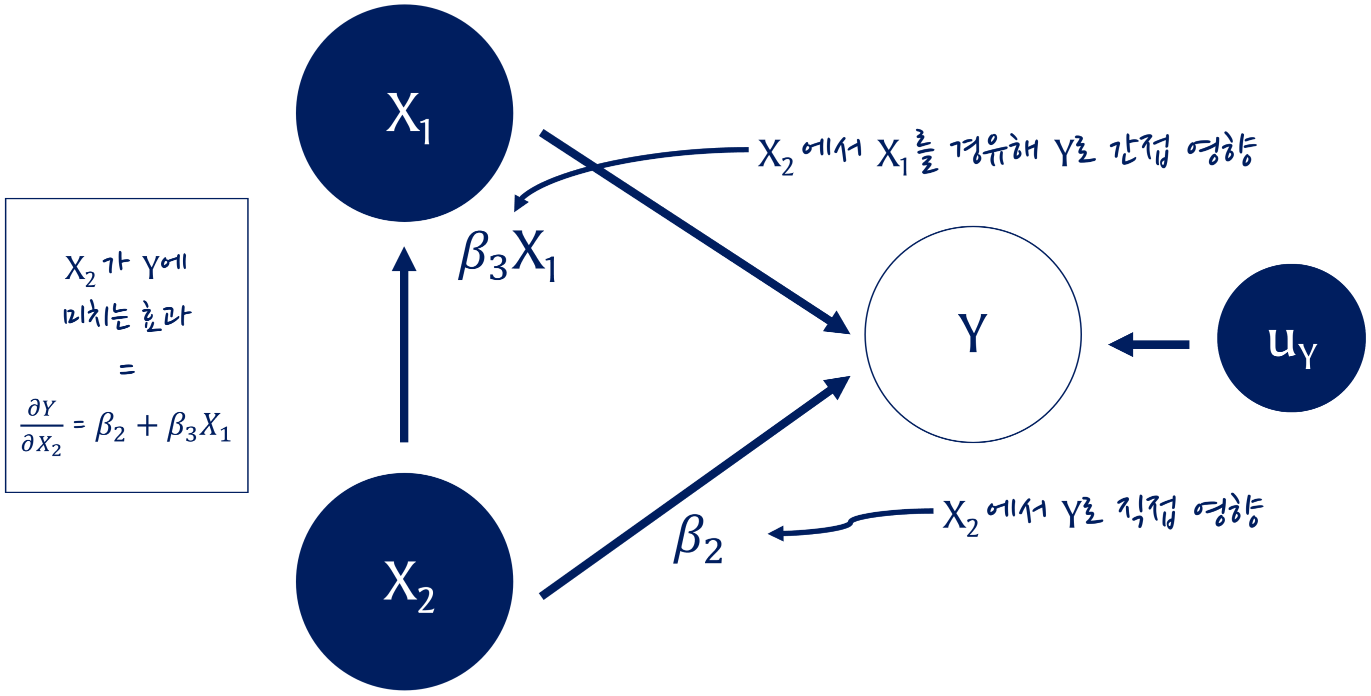
X_2가 Y에 미치는 효과 = $\frac{\partial Y}{\partial X_2} = \beta_2 + \beta_3 X_1$



다중회귀분석 II

상호작용효과: 기본모형

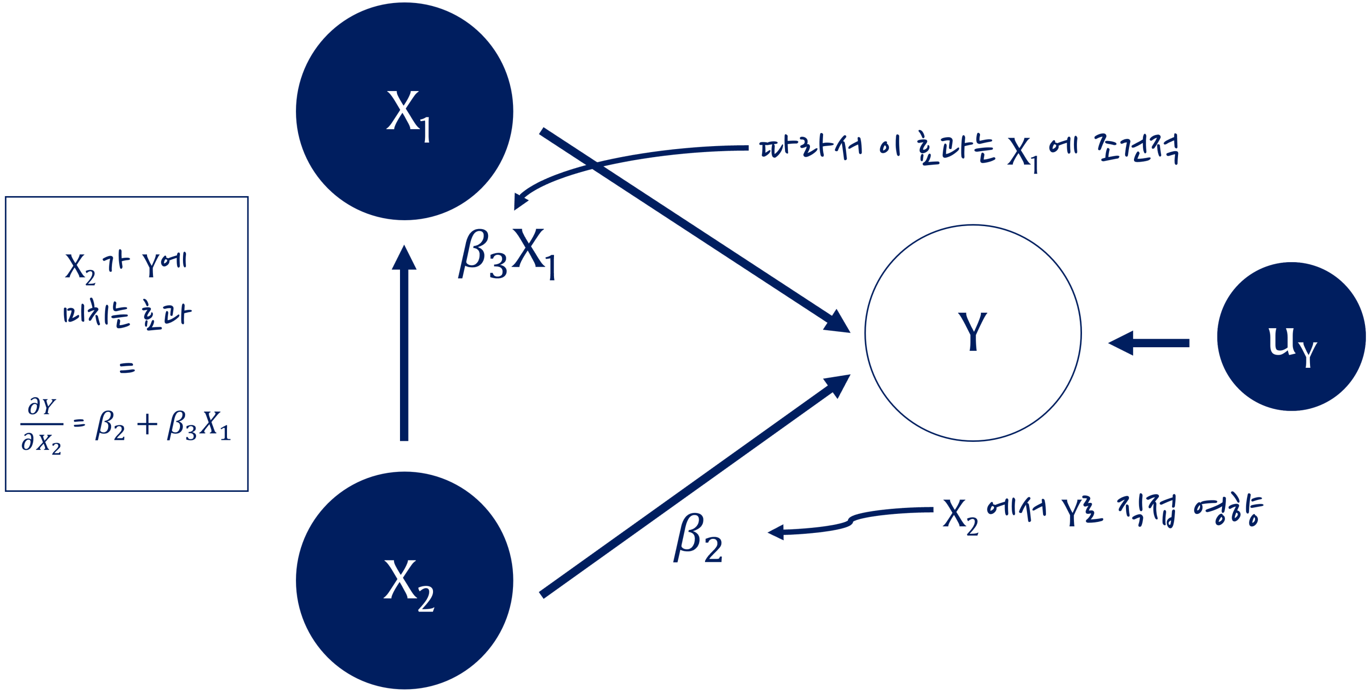
$$\boxed{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + u}$$



다중회귀분석 II

상호작용효과: 기본모형

$$\boxed{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + u}$$

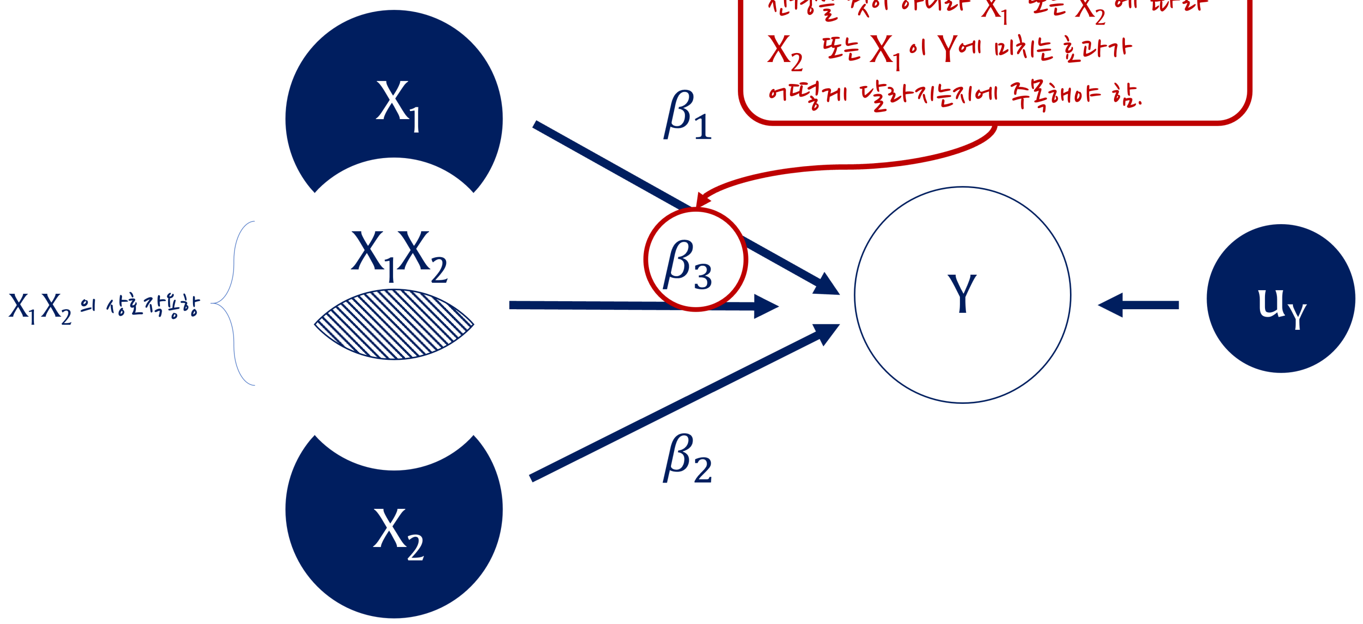


다중회귀분석 II

상호작용효과: 기본모형

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + u$$

따라서 상호작용항은 계수값의 크기에만 신경쓸 것이 아니라 X_1 또는 X_2 에 따라 X_2 또는 X_1 이 Y에 미치는 효과가 어떻게 달라지는지에 주목해야 함.



다중회귀분석 II

상호작용효과(Interaction Effects)

상호작용효과: 한 독립변수의 효과가 다른 변수의 값에 따라 달라지는 경우

- 상호작용항(interaction term): 두 독립변수 간의 곱으로 구성된 항. 하나의 설명변수 효과가 다른 변수의 수준에 따라 상호 의존적임을 모델링
- 회귀식에 상호작용항 $X \times Z$를 포함 $\rightarrow$ X의 효과가 Z에 따라 변함 (또는 Z의 효과가 X에 따라 변함)을 표현
    - 교육(X)과 경력(Z)의 상호작용: 교육의 임금효과가 경력에 따라 달라질 수 있음 (경력이 많을수록 교육의 효과가 높아진다든지 등)

앞서 하나의 예측변수와 종속변수 간의 관계를 파악하기 위해서 산포도를 그리고, 그 둘의 관계를 나타내는 경향선(trend line)을 그린 바 있음.

다중회귀분석 II

상호작용효과(Interaction Effects)

두 변수 X와 Z의 상호작용을 포함한 다중회귀모형

$$y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 (X \times Z) + u$$

이를 전개하면: $Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ + u$

이 모형에서는 X의 기울기(효과)가 Z에 의존함

- $Z = 0$일 때: $Y = \beta_0 + \beta_1 X$ (β_1는 $Z = 0$에서 X의 효과)
- $Z = 1$일 때: $Y = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)X$ (절편 $\beta_0 + \beta_2$, 기울기 $\beta_1 + \beta_3$)

즉, Z 값이 변함에 따라 X의 효과(기울기)가 β_3만큼 변동. β_3는 **" X의 효과가 Z 한 단위 증가 시 얼마나 변하는가"**를 나타냄

다중회귀분석 II

한계효과(Marginal Effects)

상호작용 모형에서 X의 한계효과(partial derivative)

- $\partial Y / \partial X = \beta_1 + \beta_3 Z$: Z의 값에 따라 X의 효과가 선형적으로 변화
- 마찬가지로 Z의 한계효과: $\partial Y / \partial Z = \beta_2 + \beta_3 X$

β_1과 β_2는 각각 $Z = 0$, $X = 0$에서의 효과를 나타내며, β_3가 0이 아니면 X와 Z가 상호작용하여 효과가 조건부로 달라짐.

- β_3 > 0이면 Z 증가에 따라 X의 효과가 증가 (양의 상호작용), β_3 < 0이면 Z 증가에 따라 X의 효과가 감소
- 그렇다면 일반 다중회귀모형 $y = \beta_0 + \beta_1 X + \beta_2 Z + u$에서 X의 Y에 대한 한계효과는 무엇인가?

다중회귀분석 II

이산형 vs 연속형 상호작용

Z가 이산형 (예: 0/1 더미): 상호작용모형은 두 집단의 회귀선을 서로 다른 기울기로 추정하는 것과 동일

- $Z = 0$ (예: 통제집단) vs $Z = 1$ (처치집단)의 기울기 비교
- 교육수준(X)의 임금효과가 남성($Z = 0$)과 여성($Z = 1$)에서 다른 경우 $\rightarrow$ β_3로 차이 포착

Z가 연속형: Z의 값에 따라 X의 효과가 연속적으로 변화

- 소득(X)의 행복도 효과가 연령(Z)에 따라 점진적으로 달라짐 (젊을 때 vs 노년기 영향 차이)
- 한계효과 $\beta_1 + \beta_3 Z$는 Z의 함수(직선) $\rightarrow$ Z의 증감에 따라 X 효과가 점진적으로 증가/감소

다중회귀분석 II

이산형 vs 연속형 상호작용

상호작용 효과 해석 시 유의사항

Brambor, Golder and Clark (2006)는 상호작용 회귀분석에서 다음 모범 규칙을 제시

모든 구성항(주효과) 포함

- 상호작용을 구성하는 개별 변수 X, Z를 모형에 반드시 포함할 것
- $X \times Z$만 포함하고 Z (또는 X)를 누락하면 심각한 모형 누락으로 계수 해석이 왜곡되고 추정량의 편향 가능
- 단, X나 Z가 자연적 0값이 없어 해석 무의미한 경우에도 주효과는 포함하고 해석은 보조적으로 다룸 (계수 해석은 기준 값에서의 효과로)

다중회귀분석 II

이산형 vs 연속형 상호작용

상호작용 효과 해석 시 유의사항

주효과 계수 단독 해석 금지: β_1이나 β_2는 특정 기준에서의 효과일 뿐, 전반적 영향 아님

- β_1은 $Z = 0$에서만 X의 효과 (만약 $Z = 0$인 사례가 현실에 없으면 β_1 자체로는 의미 없음)
- 상호작용 있으면 각 변수의 평균적(무조건적) 효과는 별도 계산 필요: 전 범위 평균 효과는 가중치에 따라 달라짐.
- $Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_2 X_3$에서 β_2를 X_2의 한계효과로 설명했으나, 실제로 β_2는 $X_3 = 0$일 때만 해당

다중회귀분석 II

이산형 vs 연속형 상호작용

상호작용 효과 해석 시 유의사항

상호작용항 유의성에만 의존하지 말 것: β_3의 p-value만 보고 상호작용 효과 유무를 섣불리 판단하면 안 됨

β_3가 유의하지 않아도, X의 효과 $\beta_1 + \beta_3 Z$가 일부 Z 값 구간에서 유의할 수 있음

반대로 β_3 유의하다고 모든 범위에서 조건부 효과가 유의한 것은 아님

상호작용 모델 채택 여부는 이론적 기대 + 조건부 효과 검토에 기반해야 하며, 단순히 β_3 $p-$값이 0.07이라 빼버리는 건 위험

다중회귀분석 II

이산형 vs 연속형 상호작용

상호작용 효과 해석 시 유의사항

항상 조건부 한계효과와 표준오차를 제시: 결과표만으로는 상호작용 효과를 제대로 전달 못함

- 관심 있는 X의 효과 $\beta_1 + \beta_3 Z$를 Z의 다양한 값에서 계산하고, 해당 신뢰구간을 함께 그래프로 보여줄 것
- Z가 이분형이면 $Z = 0, 1$에서의 X 효과와 SE 2개, Z 연속이면 효과-Z 곡선 및 유의성 구간 (영향이 유의한 Z 범위) 표시
- 표준 회귀결과표는 $Z = 0$에서의 X 효과(β_1)와 그 SE만 제공 $\rightarrow$ $Z \neq 0$의 정보는 계산 없인 알 수 없고, 특히 Z가 연속일 때 표 전체로는 파악 불가

다중회귀분석 II

이산형 vs 연속형 상호작용

예제: 교수 평가 점수 데이터(`evals`)

데이터: 미국 대학 강의평가 (evals) - 교수의 특징과 학생평가 점수

변수

- `score`: 교수 종합 평가점수 (5점 만점 평균)
- `bty_avg`: 교수의 외모 매력도 점수 (6명 평가 평균, 1~10)
- `gender`: 교수 성별 (female/male)

가설: 교수의 외모 효과가 성별에 따라 다를 수 있다(예: 남성 교수의 경우 외모가 점수에 큰 영향, 여성 교수는 상대적으로 영향 적음?)

다중회귀분석 II

이산형 vs 연속형 상호작용

예제: 교수 평가 점수 데이터(evals)

```
m_evals <- lm(score ~ bty_avg * gender, data = evals)
get_regression_table(m_evals)
```

```
## # A tibble: 4 × 7
##   term                estimate std_error statistic p_value lower_ci upper_ci
##   <chr>                  <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept              3.95      0.118     33.5    0         3.72     4.18
## 2 bty_avg                0.031     0.024      1.28   0.202    -0.017    0.078
## 3 gender: male          -0.184     0.153     -1.20   0.232    -0.485    0.118
## 4 bty_avg:gendermale     0.08      0.032      2.45   0.015     0.016    0.143
```

다중회귀분석 II

이산형 vs 연속형 상호작용

예제: 교수 평가 점수 데이터(`evals`)

위 회귀표에서 (기준범주: female)

- `bty_avg`: 여성 교수에서 외모점수 1점 증가가 $\hat{\beta}_1$만큼 평가점수 변화 (주효과)
- `gender`: male: 남성 교수와 여성 교수의 절편 차이 ($\hat{\beta}_2$) - 외모점수 0일 때 점수 차이 (외모 0은 이론적 값)
- `bty_avg:gender`: male (상호작용 $\hat{\beta}_3$): 남성 교수의 외모 기울기 증가분. 남성의 외모 영향 = 여성의 외모계수 + $\hat{\beta}_3$

다중회귀분석 II

이산형 vs 연속형 상호작용

예제: 교수 평가 점수 데이터(`evals`)

만약 결과가:

- $\hat{\beta}_1$ = 0.03: 여성 교수는 외모 1점 $\uparrow \rightarrow$ 점수 0.03 점 $\uparrow$
- $\hat{\beta}_3$ = 0.08: 남성 교수의 외모 효과는 여성보다 0.0606점 높음 → 남성의 외모 1점↑ 효과 = 0.03 + 0.0306 = 0.0606점 $\uparrow$
- $\hat{\beta}_2$ = -0.18: 성별 자체 효과는 유의하지 않을 수 (동일 외모점수에서 성별차이 적음)

즉, 남성 교수의 외모가 평가에 더 큰 영향을 미침(가설과 일치). 여성 교수의 외모 효과는 작거나 유의치 않을 수 있음.

다중회귀분석 II

조건부 한계효과 계산(margins)

{marginaleffects}로 성별별 외모 효과와 SE를 쉽게 구할 수 있음.

```r
library(marginaleffects)
marginaleffects::avg_comparisons(m_evals, by = "gender")
```

```
##
##     Term      Contrast gender Estimate Std. Error    z Pr(>|z|)    S   2.5 %
##  bty_avg +1            female   0.0306     0.0240 1.28  0.20175  2.3 -0.0164
##  bty_avg +1            male     0.1103     0.0219 5.04  < 0.001 21.1  0.0674
##  gender  male - female female   0.1874     0.0503 3.72  < 0.001 12.3  0.0888
##  gender  male - female male     0.1543     0.0505 3.06  0.00225  8.8  0.0553
##  97.5 %
##  0.0777
##  0.1531
##  0.2861
##  0.2532
##
## Type:  response
```

다중회귀분석 II

조건부 한계효과 계산(margins)

`{marginaleffects}`로 성별별 외모 효과와 SE를 쉽게 구할 수 있음.

- 위 결과에서 각 행은 성별별 `bty_avg` 한계효과
  - 여성(`gender == female`): 효과 = β_{bty}
  - 남성(`gender == male`): 효과 = $\beta_{bty} + \beta_{intercept}$

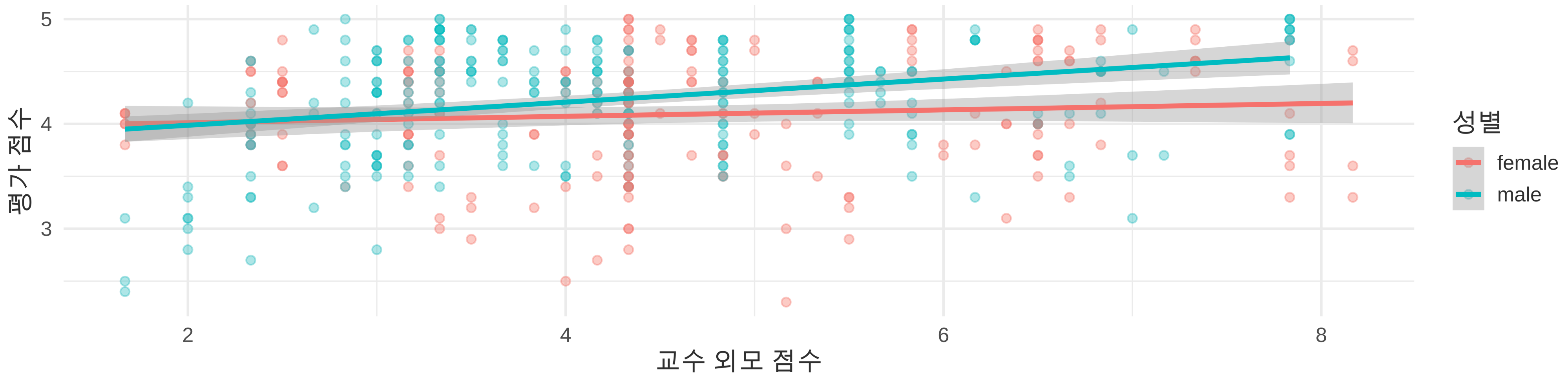
표준오차와 $p-$값도 함께 제시되어, 남성의 외모효과가 유의한지 여부 등을 직접 확인 가능

다중회귀분석 II

조건부 한계효과 계산(margins)

성별에 따른 외모점수와 평가점수 관계

```
evals |> ggplot(aes(x=bty_avg, y=score, color=gender)) +
  geom_point(alpha=0.4) +
  geom_smooth(method="lm", se=TRUE) +
  labs(x="교수 외모 점수", y="평가 점수", color="성별")
```



다중회귀분석 II

연속형 상호작용: 나이 × 외모

가정: 나이에 따라 외모의 영향이 달라질 수 있다.

- 젊은 교수 vs 나이 많은 교수에게 학생들이 외모를 평가에 반영하는 정도 차이?

다중회귀분석 II

연속형 상호작용: 나이 × 외모

회귀모형: `score ~ bty_avg * age`

```
m_evals2 <- lm(score ~ bty_avg * age, data = evals)
get_regression_table(m_evals2)
```

```
## # A tibble: 4 × 7
##   term          estimate std_error statistic p_value lower_ci upper_ci
##   <chr>            <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept         5.16     0.368     14.0    0         4.43     5.88
## 2 bty_avg          -0.188    0.076     -2.48   0.013   -0.337   -0.039
## 3 age              -0.026    0.007     -3.56   0       -0.041   -0.012
## 4 bty_avg:age       0.005    0.002      3.37   0.001    0.002    0.008
```

다중회귀분석 II

연속형 상호작용: 나이 $\times$ 외모

회귀모형: `score ~ bty_avg * age`

- `bty_avg`: 나이 0일 때(이론값) 외모 효과 β_1
- `age`: 외모점수 0일 때 나이 효과 β_2 (나이 그 자체 영향)
- `bty_avg:age`: 나이 한 살 증가에 따른 외모 효과 변화 β_3

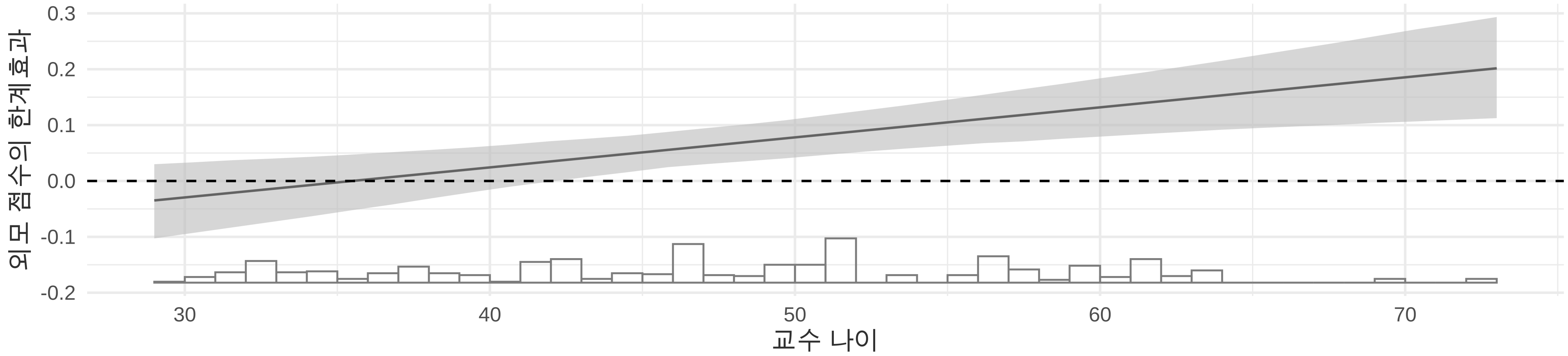
만약 $\hat{\beta}_3$가 음수라면, 연령이 높을수록 외모의 평가 영향이 감소한다는 뜻. 양수면 반대.

다중회귀분석 II

연속형 상호작용: 나이 × 외모

실제 데이터에서 β_3가 유의하지 않을 수 있음 (학생들이 나이에 관계없이 외모영향 일정)

```r
library(interplot)
interplot(m = m_evals2, var1="bty_avg", var2="age", hist=TRUE) +
  labs(x="교수 나이", y="외모 점수의 한계효과") +
  geom_hline(yintercept=0, linetype="dashed")
```



다중회귀분석 II

연속형 상호작용: 나이 × 외모

이 `interplot` 그래프는 교수 나이에 따른 외모 효과와 95% CI를 보여줌. 아래 히스토그램은 표본 내 나이 분포.

- 만약 선이 0선 위에서 시작해 나이에 따라 내려오면: 젊은 교수일수록 외모가 점수에 영향 큼, 나이 많아지면 효과 줄어들 가능성
- 신뢰구간이 0을 포함하지 않는 영역: 그 구간에서는 외모 효과 유의. 예컨대 30대~40대 구간에서는 0선 위 (유의한 양효과), 60대 이상에서는 0 근처(유의하지 않음)일 수 있음.

또한 데이터 분포를 함께 보면, 효과가 유의한 구간에 실제 관측치가 얼마나 존재하는지 판단 가능(region of significance 논의)

다중회귀분석 II

연속형 상호작용: 나이 $\times$ 외모

Hainmueller, Mummolo, and Xu (2019)의 보완점

최근 연구로 효과를 재분석한 결과, 두 가지 문제가 자주 간과됨:

- 선형 상호작용 가정: 상호작용을 직선 형태(계속 일정한 기울기 변화)로 가정하지만, 실제로는 비선형적 상호작용일 수 있음
    - 즉, Z 증가에 따라 X 효과가 일정한 속도로 변한다는 가정이 틀릴 가능성
    - 처음에는 Z 증가로 X 효과 많이 변하다가 어느 수준 이후 포화되는 등 곡선 관계
- 공통 지원 부족: 조절변수(Z) 값 범위 중 실제 데이터가 희소한 영역에서는 추정된 조건부 효과가 신뢰하기 어려움

다중회귀분석 II

연속형 상호작용: 나이 $\times$ 외모

Hainmueller, Mummolo, and Xu (2019)의 보완점

상호작용모형 추정은 Z 값 조합에 데이터가 있어야 의미있는데, **외삽(extrapolation)**으로 얻은 효과는 불안정

- Z가 매우 클 때의 X 효과를 추정했는데 표본에 그런 Z 값 거의 없음 → 그 부분 결과는 불확실

많은 논문의 상호작용 결과가 이 두 문제로 인해 모형 의존적 or 실제보다 과장/오도되었을 수 있음.

다중회귀분석 II

연속형 상호작용: 나이 $\times$ 외모

Hainmueller, Mummolo, and Xu (2019)의 보완점

간단한 진단 체크리스트와 유연한 추정 방법 제안

- 선형 가정 진단: 비선형 상호작용 여부 확인
  - Z를 구간화하거나 다항식(term)으로 확장하여 적합해보기, 또는 non-parametric 방법으로 X 효과 곡선 추정
  - 상호작용항 외에 XZ^2 같은 2차항 추가하여 곡선 형태 모형 적합, 유의성 검정 등
  - 만약 선형모형과 차이가 크면, 단순 선형 상호작용으로는 충분치 않음을 시사

다중회귀분석 II

연속형 상호작용: 나이 $\times$ 외모

Hainmueller, Mummolo, and Xu (2019)의 보완점

간단한 진단 체크리스트와 유연한 추정 방법 제안

- 공통지원 진단: 데이터 분포 확인
  - Z의 분포 및 X - Z 조합 분포 시각화 (예: 산포도나 2D 히트맵, `hist=TRUE` 옵션 등)
  - 조건부 효과 그래프에 표본 분포(histogram)를 그려, 효과가 유의한 구간에 표본이 충분히 존재하는지 보고 판단
  - 극단적인 Z값에서만 유의한 효과가 나오고 실제 그런 사례 거의 없다면, 그 효과의 실질적 중요성은 낮다고 봐야

다중회귀분석 II

연속형 상호작용: 나이 $\times$ 외모

Hainmueller, Mummolo, and Xu (2019)의 보완점

간단한 진단 체크리스트와 유연한 추정 방법 제안

- 유연한 모형 추정: 상호작용 효과의 비선형성을 허용하는 방법 도입
  - 국소적 회귀/스플라인을 활용한 `{interflex}` 패키지 (R)
  - 비선형 상호작용과 자료 부족 영역을 자동으로 처리

X의 효과를 Z의 함수로 더 유연하게 추정, 과도한 외삽 피하면서도 패턴 포착

조건부 효과 그래프에 더해, 효과가 유의한 영역의 표본비율 등을 보고, 추정에 과도한 자신감 배제 (필요 시 다중비교 보정 등)

다중회귀분석 II

연속형 상호작용: 나이 × 외모

Hainmueller, Mummolo, and Xu (2019)의 보완점

상호작용 효과를 연구할 때는 이론적 맥락에 부합하는 모형 설정과 함께, 조건부 효과를 빠짐없이 탐색하고 **모형 가정의 타당성(선형성, 공통지원)**을 확인해야 함.

이를 통해 결과의 신뢰성을 높일 수 있음.

많은 기존 연구에서도 이러한 추가검토를 거치면 결론이 달라질 수 있음을 시사.

KNU 강원대학교
KANGWON NATIONAL UNIVERSITY

감사합니다!

궁금한 것이 있으면 언제든 연락하세요.

강사 연락처

연락처	박상훈
	sh.park.poli@gmail.com
	sanghoon-park.com/
	영상바이오관 405