계량분석
다변량 회귀분석 (I)

김현우, PhD[1]

[1] 충북대학교 사회학과 조교수

November 8, 2021

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

진행 순서

1. 지난 주 리뷰
2. 이변량 회귀모형의 유의성 검정
3. 다변량 회귀모형
4. 모형의 적합도

지난 주 리뷰

지난 주 리뷰

상관분석([Stata 코드] 참고).

- 둘 이상의 숫자형 변수 사이의 관계를 볼 때는 일차적으로 상관분석(correlation analysis)을 수행한다.
- 상관계수는 반드시 -1과 1사이에 놓인다.
- 상관계수가 0보다 크면 두 변수는 서로 같은 방향(정방향)으로 움직인다. 즉 $\mathrm{Cov}(x, y) > 0$ 이면 x가 커지면 y도 커진다. 반대로 0보다 작으면...
- 상관계수가 1에 가까울수록(그리고 -1에 가까울수록) 두 변수는 더욱 밀접한 관계를 갖게 된다.
- 해석은 이런 식이 편리하다: 0과 1 사이를 사분위수로 나누고 각각 리커트 4점 척도(1사분위수=상관관계가 없다; 2사분위수=상관관계가 약하다; 3사분위수=상관관계가 어느 정도 있다; 4사분위수=상관관계가 강하다)로 의미를 부여하여 해석한다. 물론 0과 -1 사이에서도 마찬가지이다.
- 상관계수를 보고할 때는 반드시 함께 산포도를 그려보자.
- 상당히 많은 논문들이 기술통계의 일환으로 상관계수행렬을 제시한다.

충북대학교 CHUNGBUK NATIONAL UNIVERSITY

지난 주 리뷰

이변량 단순회귀분석([Stata 코드] 참고).

- 이른바 선형모형(linear model)은 아래와 같이 설정할 수 있다.

$$y_i = \beta_0 + \beta_1 X_i + u_i$$

- 일단 우리는 숫자형 척도로 측정된 종속변수와 독립변수를 사용한다.
- 데이터를 관통하는 하나의 선이 바로 회귀분석(regression analysis)의 핵심이다. 회귀모형에 따르면 오차(error)를 최소화하는 적합선이야말로 "가장 잘 맞는 직선(best-fitting straight line)"이다.
- 회귀분석의 결과표는 크게 (1) 분산분석(ANOVA), (2) 요약정보, (3) 추정된 회귀모형의 세 부분으로 나뉜다.
- 다른 추리통계학과 마찬가지로 회귀분석 역시 표본을 넘어 모집단의 성격을 추리해야 한다. t 검정을 통해 표본의 검정 통계량에 대한 유의성 검정을 수행할 수 있다.
- 특히 세 가지 적합도 지표(goodness-of-fit indices)가 중요하다: 결정계수(R^2), 조정된(adjusted) 결정계수(R^2), 그리고 RMSE.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

이변량 회귀모형의 유의성 검정

이변량 회귀모형의 유의성 검정

회귀모형을 배울 때는 이변량(bivariate)과 다변량(multivariate)을 구분한다.

- 앞서 배운 회귀모형에서 우리는 오로지 독립변수(X)와 종속변수(y), 단 두 개의 변수만 고려하였다. 종속변수 1개와 독립변수 1개가 모델 안에 투입되었으므로 이것을 이변량 회귀모형이라고 부른다.

$$y_i = \beta_0 + \beta_1 X_i + u_i$$

- 반면, (종속변수는 여전히 1개지만), 모델 안에 독립변수가 여러 개 투입된 경우를 다변량 회귀모형(multivariate regression model) 또는 다중회귀모형(multiple regression model)이라고 부를 수 있다.

$$y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki} + u_i$$

- 다변량 모형의 경우도 하첨자로 i가 붙어있으므로 관측치(observations)에 따라 y_i와 X_i, u_i가 달라지며, X_1, X_2, ..., X_k 는 모두 "다른 독립변수"를 의미한다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

이변량 회귀모형의 유의성 검정

독립변수(X)의 값이 한 단위 변화(unit change)하면 회귀계수(b_1)만큼 종속변수(y)에 영향을 미친다(**[Stata 코드]** 참고).

- 회귀계수 및 상수의 해석은 무척 단순하지만 연습을 필요로 한다.
- 가령 결석 시수(skipped)가 올해 학점(termgpa)에 영향을 미친다는 회귀모형을 세우고 "오차항을 최소화하는 b_0와 b_1"을 다음과 같이 추정하였다고 하자.

$$\text{termgpa}_i = 3.043 - 0.076\text{skipped}_i$$

- "결석 시수가 한 단위 증가할 때, 올해 학점은 -0.076점만큼 감소한다."
- "결석 시수가 0일 때의 올해 학점은 3.043이다."

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

이변량 회귀모형의 유의성 검정

추정된 상수($\widehat{b_0}$)와 회귀계수($\widehat{b_1}$)를 가지고 예측된 Y(predicted Y; $\hat{y}$)를 구할 수 있다(**[Stata 코드]** 참고).

- 추정된 변수에 대해서는 이렇게 hat을 붙인다.
- 우리가 추정한 모형에서 $\widehat{b_0} = 3.043$ 이고 $\widehat{b_1} = -0.076$이므로 독립변수인 skipped_i에 원하는 값을 대입하면 종속변수인 termgpa를 예측(prediction)할 수 있다.

$$\widehat{\text{termgpa}_i} = 3.043 - 0.076\text{skipped}_i$$

- Stata에서는 **predict** 명령어로 주어진 관찰값의 특성에 따른 $\hat{y}$을 구할 수 있다.
- 정의상 오차(error)는 $e_i = y - \hat{y}$이므로 표본 안에서의 오차도 구해볼 수 있다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

이변량 회귀모형의 유의성 검정

회귀계수와 상수에 대해서 (t 분포를 사용한) 유의성 검정을 수행한다.

- 우리는 표본을 가지고 회귀분석을 수행하여 b_0과 b_1을 추정하였다. 그러나 진짜 궁금한 것은 모집단에서의 β_0과 β_1를 아는 것이다. 이제부터 회귀계수(b_1)에 대한 유의성 검정 절차를 살펴보자.
- 회귀계수에 관한 영가설은 $H_0 : \beta_1 = 0$ 이고 대립가설은 $H_a : \beta_1 \neq 0$이다.
- 표본을 무한히 추출하여 각각으로부터 계속 회귀분석을 수행하여 $\hat{b}_1$을 무한히 추정한다고 하자. (오차항에 대해 몇 가지 가정을 하면) 그 상수와 회귀계수의 표집분포는 소표본일 때 t 분포를, 대표본일 때 정규분포한다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

이변량 회귀모형의 유의성 검정

- 우리는 회귀분석에 관한 몇 가지 가정이 충족된다는 전제 아래, 중심극한정리에 힘입어 다음을 알 수 있다.

$$\beta_1 = \mathrm{E}(\widehat{\mathrm{b}}_1)$$

- 한편 표준오차란 표집분포의 표준편차를 의미한다. 그러므로 회귀계수의 표준오차(standard error)는 추정값(estimates)에 대해 얼마나 확신하는가(confident)를 보여준다.

$$\mathrm{SE}(\widehat{\mathrm{b}}_1) = \sqrt{\frac{\sum(\mathrm{y_i} - \widehat{\mathrm{y}}_\mathrm{i})^2/(\mathrm{n}-\mathrm{k}-1)}{\sum(\mathrm{x_i} - \bar{\mathrm{x}})^2}} = \sqrt{\frac{\mathrm{SS_{error}}/(\mathrm{n}-\mathrm{k}-1)}{\mathrm{SS_{total}}}}$$

- k는 "상수를 제외한 패러미터(parameter)", 즉 독립변수의 수이다. 이변량의 경우 $\mathrm{k} = 1$이다. 분산분석의 용어를 빌리면 분자는 $\mathrm{MS_{error}}$ 라고 불릴수 있고, 분모는 독립변수의 $\mathrm{SS_{total}}$ 이라고 불릴 수 있다(Why?).

충북대학교 CHUNGBUK NATIONAL UNIVERSITY

이변량 회귀모형의 유의성 검정

- 이제 우리에게 주어진 표본을 가지고 회귀분석을 수행하여 얻은 회귀계수 추정값(estimates), 즉 $\hat{b}_1$ 의 t 값을 다음과 같이 계산할 수 있다. 그런데 $\beta_1 = 0$ 이므로 (Why?) 식은 조금 더 단순하게 고쳐 쓸 수 있다.

$$t_{\hat{b}_1} = \frac{\hat{b}_1 - \beta_1}{\widehat{SE}_{\hat{b}_1}} = \frac{\hat{b}_1}{\widehat{SE}_{\hat{b}_1}}$$

- 위에서 제시한 영가설이 옳다는 전제 아래 그린 표집분포에 따랐을 때, 표본에서 추정된 $t_{\hat{b}_1}$ 의 위치를 확인해보고 그보다 극단적인 $t_{\hat{b}_1}$ 값을 얻게 될 확률, 즉 유의확률(p-value)을 계산할 수 있다.
- 만일 유의확률이 0.05보다 작다면 우리는 5% 유의수준 또는 95% 신뢰수준에서 영가설($H_0 : \beta_1 = 0$)을 기각하고 대립가설($H_a : \beta_1 \neq 0$)을 채택할 수 있다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

이변량 회귀모형의 유의성 검정

Stata에서 **regress** 명령어로 회귀분석을 수행하면 유의성 검정의 결과를 확인할 수 있다(**[Stata 코드]** 참고).

- 아래 표에서 회귀계수(Coef.), 표준오차(Std. Err.), t 값(t), 유의확률(P>|t|), 95% 신뢰수준(95% Conf. Interval)을 꼼꼼히 확인해보자.

```
. reg termgpa skipped
```

Source	SS	df	MS	Number of obs	=	680
				F(1, 678)	=	309.40
Model	115.436802	1	115.436802	Prob > F	=	0.0000
Residual	252.960735	678	.373098429	R-squared	=	0.3133
				Adj R-squared	=	0.3123
Total	368.397537	679	.542558964	Root MSE	=	.61082

termgpa	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
skipped	-.0755857	.0042971	-17.59	0.000	-.084023	-.0671484
_cons	3.043399	.0343692	88.55	0.000	2.975916	3.110881

충북대학교 CHUNGBUK NATIONAL UNIVERSITY

이변량 회귀모형의 유의성 검정

보다 능동적으로 $H_0 : \beta = x$라는 영가설을 텍스트해 볼 수도 있다(**[Stata 코드]** 참고).

- 유의성 검정에서 영가설은 $H_0 : \beta = 0$였으므로 이론 확인에는 너무 제한적이라고 느꼈을수도 있다. 그러면 사후검정(post-estimation)이 필요하다.
- 이른바 왈드 검정(Wald test)라고 불리우는 가설검정 기법은 다음과 같은 가설구조를 F 검정으로 테스트할 수 있다. 물론 β_1 하나

$$H_0 : \beta_1 = \beta_2 = ... = \beta_k = x$$
$$H_a : (\beta_1 \neq \beta_2) \text{ or } (\beta_2 \neq \beta_3) \text{ or } ... \text{ or } (\beta_{k-1} \neq \beta_k)$$

- 물론 하나의 β 가 특정 값을 가지는가에 관해 가설검정을 할 수 있다. 특정 조합, 예를 들면 $(\beta_1 = m)\&(\beta_2 = n)$에 대해서도 가설검정을 할 수도 있다.

$$H_0 : (\beta_1 = m) \text{ and } (\beta_2 = n)$$
$$H_a : (\beta_1 \neq m) \text{ or } (\beta_2 \neq n)$$

충북대학교 CHUNGBUK NATIONAL UNIVERSITY

이변량 회귀모형의 유의성 검정

통계적 유의성(statistical significance)와 실질적 유의성(substantial significance)은 다른 개념이다

- 통계적 유의성을 해석할 때 가장 흔한 실수 중 하나는 유의확률(p-value)이 작은 것을 가지고 선형적 관계의 강도(strength of linear relationship)로 해석하는 일이다. 유의확률은 단지 $H_0 : \beta = 0$라는 옳은 영가설을 기각하는 가능성을 보여줄 뿐이다.
- 실질적 유의성은 통계적으로 유의한가 여부와 상관없이 실제로 얼마나 그 강도가 센가의 문제를 다룬다. 예컨대 한 시간 게임을 더하게 되면 독서시간이 2분 줄어든다는 발견(Cummings and Vandewater 2007)은 설령 통계적으로 유의하더라도 실질적으로는 그다지 유의하지 않다.
- 그러므로 통계적으로 유의한 결과를 얻었더라도 그 관계의 그래프를 그려보고 실제로 해석해보아 실질적 유의성이 얼마나 높은지 판단할 필요가 있다.

Cummings, Hope M. and Elizabeth A. Vandewater. 2007. "Relation of Adolescent Video Game Play to Time Spent in Other Activities." Archives of Pediatrics Adolescent Medicine 161(7): 684-689.

충북대학교 CHUNGBUK NATIONAL UNIVERSITY

이변량 회귀모형의 유의성 검정

분산분석표(ANOVA table)는 일원분산분석 결과를 통해 모형의 적합도(goodness-of-fit)를 보여준다(**[Stata 코드]** 참고).

- 이변량 회귀분석에서 사용된 일원분산분석의 영가설은 $H_0 : \beta_1 = 0$이다.
- 일원분산분석의 검정통계량 F 값은 다음과 같다. 단 k는 독립변수의 수를 의미한다.

$$F_{(k,n-k-1)} = \frac{MS_{between}}{MS_{within}} = \frac{SS_{between}/k}{SS_{within}/(n-k-1)}$$

- F 값이 클수록 유의확률이 작아져(Why?) 위 영가설을 기각할 수 있다.
- 이변량 회귀분석에서는 영가설이 t 검정과 같기 때문에 (앞서 배운) $F = t^2$ 의 공식이 성립한다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

다변량 회귀모형

다변량 회귀모형

다변량 회귀분석은 (다른 변수들의 효과를 통제한 상태에서) 특정 변수의 순효과(net effect) 또는 부분효과(partial effect)를 살펴보는데 유용하다.

- 여러 개의 독립변수를 모델에 투입했다면 여러 영향력은 각각에 해당되는 변수 안으로 분산 흡수된다. 그러므로 다변량 회귀분석은 다른 변수의 영향력으로부터 독립된 특정 변수의 영향력을 살펴보는데 유리하다.
- 예컨대, 로그 평준화임금(ln_wage)을 종속변수로, 나이(age) 및 직무경험(ttl_exp)을 독립변수로 하는 회귀모형을 통해 "나이(age)의 효과를 통제했을 때(즉 같은 나이일 때)" 직무경력(ttl_exp)이 한 단위 변화하면 임금(ln_wage)이 얼마만큼 변화하는지 파악할 수 있다.

$$\text{ln\_wage} = b_0 + b_1 \text{age} + b_2 \text{ttl\_exp} + e$$

- 이렇게 "다른 요소들이 같다고 할 때(all other things being equal)"라는 표현에 대하여 경제학에서는 ceteris paribus 라는 라틴어를 사용하기도 한다."

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

다변량 회귀모형

다변량 회귀분석의 수행에서는 변수 체크에 주의를 기울여야 한다(**[Stata 코드]** 참고).

- 여러 개의 독립변수를 모델에 투입하다보면 하나하나를 꼼꼼하게 살펴보지 않고 그냥 대충 집어넣는 경우가 많다. 개별 변수의 척도가 어떻게 구성되어 있는지, 분포는 어떠한지, 결측치(missing values)가 있는지 등을 반드시 꼼꼼하게 살펴보아야 한다.
- Stata에서 **summarize**, **tabulate**, **codebook**, **inspect**, **histogram**, **graph** 등의 명령어를 활용해 변수를 살펴본다. 또 **edit**로 여러 변수들이 각각 어떤 구조 속에 있는지도 감을 얻어야 한다.
- 특정한 변수 별로 정렬(sort)하여 데이터를 살펴보는 것도 유용하다. Stata에서 **sort**는 그런 기능을 제공한다.
- 하나의 변수에라도 결측치가 있다면 그 행(row)을 아예 모두 삭제하자. 이것을 listwise deletion이라고 부른다. 이것은 문제를 일으킬 수 있지만 당장은 일단 이렇게 넘어간다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

다변량 회귀모형

다른 변수들의 영향력을 통제하였을 때, 특정 독립변수(X)의 값이 한 단위 변화하면 회귀계수(b)만큼 종속변수(y)에 영향을 미친다(**[Stata 코드]** 참고).

- 가령 연령(age)과 직무경험(ttl_exp)이 임금(ln_wage)에 영향을 미친다는 회귀모형을 세우고 "오차항을 최소화하는 b_0, b_1, 와 b_2"를 다음과 같이 추정하였다고 하자.

$$\text{ln\_wage}_i = 1.236 - 0.005\text{age}_i + 0.054\text{ttl\_exp}_i$$

- "직무경험의 효과를 통제하였을 때, 연령이 한 살 증가하면 로그 평준화임금은 0.005만큼 감소한다."
- "연령의 효과를 통제하였을 때, 직무경험이 한 단위 증가하면 로그 평준화임금은 0.054만큼 증가한다."
- "연령과 직무경험이 모두 0일 때의 로그 평준화임금은 1.236이다."

충북대학교 CHUNGBUK NATIONAL UNIVERSITY

다변량 회귀모형

다변량 분석에서는 "다른 변수를 통제한다는 것"의 의미를 명확하게 해야 한다(**[Stata 코드]** 참고).

- 한편 **predict** 명령어로 구한 $\hat{y}$는 개별 관찰값의 속성에 의지하므로 말하자면 대푯값은 아니다.
- 어떤 변수를 대표하는 값으로는 보통 평균(mean)이 거론되기 때문에 "통제변수의 영향력을 통제하기 위해" 관심변수를 제외한 나머지 통제변수들은 평균에 그 값을 고정(fix)시키고 해석할 수 있다.
- 만일 직무경험(ttl_exp)이 관심변수라면 연령을 평균에 고정시키고($age_i = \overline{age}$), 직무경험(ttl_exp_i)의 변화에 따른 임금(ln_wage)의 차이를 예측할 수 있다.

$$\begin{aligned} ln\_wage_i &= 1.236 - 0.005\overline{age} + 0.054 ttl\_exp_i \\ &= 1.236 - 0.005 \cdot 23.945 + 0.054 ttl\_exp_i \\ &= 1.116 + 0.054 ttl\_exp_i \end{aligned}$$

- 일단 방정식을 단지 두 변수 사이의 관계로 축약했다면 그래프도 그릴 수 있다.

충북대학교 CHUNGBUK NATIONAL UNIVERSITY

다변량 회귀모형

- 일단 종속변수에 영향을 미치는 모든 독립변수가 온전하게 측정되어 모델 안에 투입되었다고 가정하자. 이 가정을 완전모형설정(perfect model specification)이라고 부른다.
- 이 경우 종속변수에 영향을 미치는 "어떤 요인"이 변화하면 이를 측정하고 있는 변수가 이를 반영하므로, 우리는 완전히 그 "어떤 요인"이 종속변수에 대해 갖는 효과를 해당 독립변수 안으로 고립시킬 수 있다(isolate). 고립이 완벽하다면 다른 독립변수의 회귀계수에는 아무런 영향을 주지 않는다(물론 현실적으로는 다소간 영향을 주기 마련이다).
- 만일 (가정과는 달리) 그 변화를 측정하는 변수가 모델 안에 적절히 고려되지 않았다면, 그 "특정 요인"의 변화는 (1) "다른 변수"를 통해서 혹은 (2) 오차항(error term)을 통해서 종속변수에 영향을 미치게 된다.
- 그렇다면 그 "다른 변수"의 종속변수에 대한 순효과는 어느 정도 오염되었다고 할 수 있다. 우리는 그 오염된 정도를 누락변수 편의(omitted variable bias; OVB)라고 부른다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

다변량 회귀모형

우리는 종종 관심변수(variables of interest)와 통제변수(control variables)를 구별한다.

- 우리의 연구문제에 따라 "교육수준이 임금에 미치는 영향"에 관심이 있다면 교육수준이 우리의 관심변수가 되고, 나머지 (잠재적으로 임금에 영향을 미칠 수 있는) 다른 모든 독립변수들은 이른바 통제변수로 취급된다.
- 하지만 이것은 철저하게 주관적으로 붙이는 의미이자 이름일 뿐이고 모델 안에서 변수들의 지위는 완벽히 평등하다.
- 실무나 논문에서는 통제변수가 갖는 실질적 함의에 대해서 거의 논의하지 않는다. 하지만 기존 문헌에서 중요하게 보고되었던 독립변수가 빠지면 심사자에게 지적을 받기 쉽다. 그러므로 "(이를 고려해보았지만) 연구의 발견에는 큰 변화가 없었다" 하는 식으로 보고해야 할 수도 있다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

다변량 회귀모형

독립변수는 (기존문헌에 기반한) 이론적 근거와 합리적 의심에 따라 선정하여야 한다.

- 어떤 통제변수들을 고려해야 하는가? 얼마나 많은 통제변수들을 고려해야 하는가? 필요한 만큼!
- 때로는 명백히 종속변수에 영향을 미치리라고 예상되는 요인이 있지만, 그것을 측정할 수 없는 경우(unobservables)도 있을 수 있다.
- 게다가 누락변수 편의(OVB)는 그럴듯하게 몇 개의 통제변수를 고려하면 자동적으로 해소되는 그런 간단한 문제가 아니다.
- 사회현상의 복잡성을 고려한다면 완전모형설정은 사실상 불가능하다. 그러므로 누락변수 편의는 언제나 사회과학 연구의 타당성을 위협한다. 우리는 기존문헌을 꼼꼼하게 살피고 이론적 통찰력에 힘입어 변수를 구성해야 한다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

다변량 회귀모형

위계적 회귀분석(hierarchical regression analysis)을 통해 회귀계수 및 유의성 여부의 변화를 관찰할 수 있다(**[Stata 코드]** 참고).

- 단계적으로 독립변수를 모델에 투입하여 점차 회귀계수와 유의성이 변화하는 모습을 관찰할 수 있다. 이때 어떤 변수를 어떤 순서대로 넣는가는 전적으로 연구자의 몫이다.
- 아무래도 한 번에 수많은 독립변수를 넣어 근사한 회귀모형을 만들고 싶은 유혹에 빠지기 쉽지만, 하나하나 변수를 넣어가면서 차분히 회귀모형을 설계하는 것이 바람직하다. 새로 투입하는 변수을 꼼꼼히 살펴보기 위해 충분한 시간을 들여야 한다.
- 임금(ln_wage)을 설명하기 위해 (1) 나이(age)을 독립변수로 한 회귀분석을 수행해보자. 뒤이어 (2) 나이와 직무경험(ttl_exp)을 넣은 회귀분석을 수행해보자. 마지막으로 (3) 나이, 직무경험(ttl_exp), 교육년수(grade)을 넣은 회귀분석을 수행해보자. 계수와 유의성은 어떻게 변화하는가? 왜 그런가?

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

다변량 회귀모형

- 논문에서 최종적으로 보고하지 않더라도 수많은 대안모형들을 만들어 살펴보고 변수 투입에 따른 민감성을 꼼꼼히 살펴보아야 한다.
- 다변량 회귀분석에서 "거의 같은 두 개 이상의 독립변수를 동시에 집어넣으면" 이른바 다중공선성(multicollinearity) 문제가 발생한다. 예컨대 횡단면분석 (cross-sectional analysis)에서 태어난 해(birth year)와 나이(age)는 사실상 거의 같은 변수인 셈이므로(Why?) 문제를 일으키게 된다. 이에 관해서는 나중에 좀 더 자세히 학습한다.
- 참고로 위계적 회귀분석(hierarchical regression analysis)과 위계적 선형모형 (hierarchical linear model)은 상이한 것이다. 위계적 선형모형은 이른바 다층모형 (multilevel model)을 의미하며 지역/조직/시간 속의 개체를 분석할 때 유용하다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

다변량 회귀모형

Stata에서는 **esttab** 또는 **estout**으로 결과표를 쉽게 꾸밀 수 있다(**[Stata 코드]** 참고).

- 이것들은 사용자들이 만들어 배포한 명령어(user-written commands)이지만 굉장히 폭넓게 쓰인다.
- 두 명령어 모두 같은 사람이 만들었는데 접근 방식이 살짝 다르다. 상황에 따라 편리한 쪽을 사용하면 된다. 일단 한 번 만들어두면 앞으로 계속해서 쓸 수 있으므로 억지로 외우거나 하지 않아도 된다.
- **outreg2** 명령어도 있고 이쪽도 제법 유명하다. 다만 사용법이 **esttab** 또는 **estout**과는 다르다.
- 공식 명령어로는 **putexcel**나 **putdocx** 등이 있지만 이미 사람들은 **esttab**과 **estout**에 익숙해져 있어 인기가 없는 듯 하다. 게다가 공식 명령어 쪽이 사용하기 훨씬 복잡하다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

다변량 회귀모형

회귀계수를 표준화(standardization)하여 변수의 상대적 영향력을 판단할 수 있다(**[Stata 코드]** 참고).

- 어떤 독립변수가 종속변수를 가장 잘 설명하나? 회귀계수의 크기만 보고 이것을 판단할 수 없다. 이는 각 독립변수의 표준편차가 다르기 때문이다. 어떤 변수는 겨우 [1,10] 사이에 놓인 반면, 어떤 변수는 [-9999,9999] 사이에서 변화할 수 있다.
- 그러므로 비교를 위해서는 회귀계수(b)를 표준화하여 베타(β)로 표현해야 한다. 몇 가지 표준화 방식이 있는데, 가장 널리 사용되는 것은 원래 회귀계수(raw coefficient)에 독립변수/종속변수의 표준편차 비율(S_X/S_y)를 곱하는 방식인 전체 표준화 베타계수(fully standardized beta coefficients)이다.

$$\beta = b\frac{S_X}{S_y}$$

- 이제 추정된 베타계수의 크기를 보고 곧바로 여러 독립변수들 사이에서 상대적 영향력을 비교할 수 있다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

다변량 회귀모형

- 베타계수(β)의 해석은 조금 특별하다: "독립변수(X)가 1 표준편차만큼 변화할 때, 종속변수(y)는 β 표준편차만큼 변화한다."
- 원래 회귀계수(b)의 경우 "독립변수(X)가 한 단위(unit) 변화할 때, 종속변수(y)는 b 만큼 변화한다"로 해석했었다.
- 베타계수에서 사용되는 "단위"는 "표준편차"가 되는 셈이다(Why?).
- 사회학 분야의 권위있는 저널에서 살펴보면 표준화된 베타계수는 예전에 비해 그다지 인기가 없는 듯 하다.
- 실용적 목적과는 별개로 베타계수는 회귀계수가 사실 "표준화된 상관계수"에 불과하다는 점을 수학적 증명없이도 보여준다는 점에서 유용하다. 관찰자료를 가지고 평범하게 회귀분석을 수행해서 얻은 회귀계수는 결코 인과관계를 드러내 보이지 못한다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

모형의 적합도

모형의 적합도

분산분석표(ANOVA table)에서 도출된 F 값은 모형 전체의 적합도(goodness-of-fit)와 관련된다(**[Stata 코드]** 참고).

- 다변량 회귀분석 맥락에서 수행된 ANOVA의 가설구조는 다음과 같다.

$$\mathrm{H}_0 : \beta_1 = \beta_2 = \ldots = \beta_\mathrm{k} = 0$$
$$\mathrm{H}_\mathrm{a} : (\beta_1 \neq 0) \text{ or } (\beta_2 \neq 0) \text{ or } \ldots \text{ or } (\beta_\mathrm{k} \neq 0)$$

- 이 영가설을 기각하지 못한다면 설정한 모형은 완전히 무의미한 것이다. 만일 이 영가설을 기각할 수 없다면 "이 모델은 완전히 쓸모없다" 라는 영가설을 기각할 수 없는 것이나 마찬가지이기 때문이다.
- 당연히 이 가설은 Stata에서 **test**를 활용한 사후검정(post-estimation)을 통해서 재확인해 볼 수 있다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

모형의 적합도

결정계수(R^2)는 전체 변량(total variation) 대비 다변량 회귀모형에 의해 설명된 변량(explained variation)의 비율을 보여준다(**[Stata 코드]** 참고).

- 우리는 지난 주에 R^2가 모형의 높은 설명력 내지 데이터 적합도를 의미함을 배웠다.

$$R^2 = \frac{\text{Variation}_{\text{Explained}}}{\text{Variation}_{\text{Total}}}$$

- 이변량 회귀분석의 맥락에서 R^2의 제곱근($\sqrt{R^2}$)은 상관계수(R)와 수학적으로 동일하다.
- 논문에 따라서는 위계적 회귀분석을 수행한 다음, 계수를 보고하는 대신 변화하는 R^2 값만을 보고하면서 이것의 유의성 검정을 수행하는 경우도 있다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

모형의 적합도

- 교과서에 따라서는 다변량 회귀분석에서는 보정된 결정계수(adjusted R^2)를 보고하도록 권고한다.

$$\text{adjusted } R^2 = R^2 - \frac{k(1-R^2)}{n-k-1}$$

(k는 독립변수의 수)

- 사실 표본을 사용하여 추정된 R^2는 모집단을 사용하여 계산된 R^2보다 살짝 크다. 즉 R^2는 표본을 넘어 모집단에서의 모델 설명력을 제대로 보여주지 않는다. 반면 보정된 결정계수는 이를 수정한다.
- 게다가 변수를 더 많이 사용할수록 두 결정계수 사이의 차이는 점점 커진다는 점이 알려져 있다. 그러므로 다변량 회귀모형 안에 변수를 많이 사용한다면 보정된 결정계수의 사용이 더 바람직할 수도 있다.
- 실험삼아 이론적으로 쓸데없는 변수(예컨대 난수)도 독립변수로 넣어보자. **R-squared**는 조금 더 증가하지만 **Adj R-squared**는 호락호락하게 증가하지 않는다 (심지어 감소하기도 한다).

충북대학교
CHUNGBUK NATIONAL UNIVERSITY