

계량경제학 8.1

제 8 강

다중회귀모형 - 2 The Multiple Regression Model - II

계량경제학 8.4

$y_t = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} + \beta_4 x_{t4} + \varepsilon_t$

$H_0: 3\beta_2 - 7\beta_3 = 21$

$H_1: 3\beta_2 - 7\beta_3 \neq 21$

$$t = \frac{3b_2 - 7b_3 - 21}{\text{Se}(3b_2 - 7b_3)} \sim t_{(T-4)}$$

$\text{Var}(3b_2 - 7b_3) = 3^2 \text{Var}(b_2) + 7^2 \text{Var}(b_3) - 2 \times 3 \times 7 \text{Cov}(b_2, b_3)$

$$\hat{\sigma}^2 = \frac{\sum \hat{\varepsilon}_t^2}{T - K}$$

Se()

공분산 행렬 (variance-covariance matrix)을 이용

계량경제학 8.2

$y_t = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} + \beta_4 x_{t4} + \varepsilon_t$

t-검정은 모수들의 임의의 선형결합에 대한 가설에 대해서도 사용될 수 있음

$H_0: \beta_1 = 0$

$H_0: \beta_2 + \beta_3 + \beta_4 = 1$

$H_0: 3\beta_2 - 7\beta_3 = 21$

$H_0: \beta_2 - \beta_3 \leq 5$

이러한 t-검정들은 모두 정확히 T-K 자유도를 가짐
 단 K = # coefficients estimated (including the intercept).

계량경제학 8.5

F-검정

F-분포

If $V_1 \sim \chi^2_{(m_1)}$ and $V_2 \sim \chi^2_{(m_2)}$ and if V_1 and V_2 are independent, then

$$F = \frac{V_1/m_1}{V_2/m_2} \sim F_{(m_1, m_2)}$$

확률변수 F는 m_1 분자 자유도 (numerator degrees of freedom: df_n)와 m_2 분모 자유도 (denominator degrees of freedom: df_d)를 가진 F분포를 한다고 말할 수 있다.

이 확률변수는 $(0, \infty)$ 의 구간에서 밀도를 가지며 긴 오른쪽 꼬리를 가지는 모양임

계량경제학 8.3

$y_t = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} + \dots + \beta_K x_{tK} + \varepsilon_t$

In general, $H_0: \sum c_i \beta_i = c_0$

$$t = \frac{\sum c_i b_i - c_0}{\text{Se}(\sum c_i b_i)} \sim t_{(T-K)}$$

One Tail Test $H_1: \sum c_i \beta_i > c_0$

Two Tail Test $H_1: \sum c_i \beta_i \neq c_0$

계량경제학 8.6

F-검정

F-분포

$$F = \frac{V_1/m_1}{V_2/m_2} \sim F_{(m_1, m_2)} \Rightarrow \frac{1}{F} = \frac{V_2/m_2}{V_1/m_1} \sim F_{(m_2, m_1)}$$

$$t = \frac{Z}{\sqrt{V/m}} \Rightarrow t^2 = \frac{Z^2}{V/m} = \frac{V_1/m_1}{V_2/m_2} \sim F_{(1, m)}, V_1 \equiv Z^2$$

F-검정 계량경제학 8.7

일련의 가설들에 대한 F 검정은 제약이 없는(unrestricted) 원래의 모형으로부터의 잔차의 제곱의 합과 귀무가설이 참이라는 가정하의 모형으로부터의 잔차의 제곱의 합을 비교하는 것에 기반을 둔 검정방법임

F-검정 계량경제학 8.10

- 귀무가설하에서, 확률변수 F 는 분자의 자유도 J , 분모의 자유도 $T-K$ 인 F 분포를 함
- 대립가설하에서는, SSE_R 과 SSE_U 의 차이가 커지게 됨 (Why?).

F 검정통계량의 값이 너무 크게 되면 귀무가설을 기각하게 됨

F 값과 분자의 자유도 J 분모의 자유도 $T-K$ 인 F 분포의 오른쪽 꼬리의 확률을 α 로 남기는 임계값 F_c 를 비교하여 기각여부를 판단함

F-검정 계량경제학 8.8

- 귀무가설이 참이라는 가정하의 모형으로부터의 잔차의 제곱의 합:
the restricted sum of squared errors, or SSE_R
- 원래의 제약이 없는 모형으로부터의 잔차의 제곱의 합:
the unrestricted sum of squared errors, or SSE_U

$\Rightarrow SSE_R - SSE_U \geq 0$: Always True. Why?

F-검정 계량경제학 8.11

이러한 방식 F-검정에서는 귀무가설로부터의 이탈은 항상 F값을 크게 만들기 때문에 양측검정이든 단측검정이든 항상 오른쪽 꼬리 검정임

$$F = \frac{(SSE_R - SSE_U)/J}{SSE_U/(T-K)}$$

F-검정 계량경제학 8.9

- J 가 가설들의 수(즉 제약들의 수)라고 하면 다음이 성립함으로 보일 수 있음

$$V_1 = \frac{SSE_R - SSE_U}{\sigma^2} \sim \chi^2_{(J)}, \text{ under the null}$$

- 다음의 사실은 이미 살펴본 바 있음.

$$V_2 = \frac{SSE_U}{\sigma^2} \sim \chi^2_{(T-K)}$$

- 또한 V_1 과 V_2 가 통계적이 독립임도 보일 수 있음

$$F = \frac{V_1/J}{V_2/(T-K)} = \frac{(SSE_R - SSE_U)/J}{SSE_U/(T-K)} \sim F_{(J, T-K)}, \text{ under the null}$$

F-검정 계량경제학 8.12

실례 - 하나의 제약

	coefficient	std. error	t-ratio	p-value
const	104.786	6.48272	16.16	2.84e-021 ***
p	-6.64193	3.19119	-2.081	0.0427 **
a	2.98430	0.166936	17.88	4.11e-023 ***

Mean dependent var 120.3731 S.D. dependent var 16.31873
Sum squared resid 1805.168 S.E. of regression 6.069611
R-squared 0.867085 Adjusted R-squared 0.861660
F(2, 49) 159.8280 P-value(F) 3.37e-22
Log-likelihood -166.0111 Akaike criterion 338.0222
Schwarz criterion 343.8759 Hannan-Quinn 340.2664

F-검정 계량경제학 8.13

실례 - 하나의 제약

	coefficient	std. error	t-ratio	p-value
const	91.8306	1.87131	49.07	5.80e-044 ***
a	2.94907	0.171520	17.19	1.24e-022 ***

Mean dependent var 120.3231 S.D. dependent var 16.31873
 Sum squared resid 1964.758 S.E. of regression 6.268585
 R-squared 0.855334 Adjusted R-squared 0.852441
 F(1, 50) 295.6241 F-value(F) 1.24e-22
 Log-likelihood -168.2137 Akaike criterion 340.4274
 Schwarz criterion 344.3299 Hannan-Quinn 341.9235

F-검정 계량경제학 8.16

실례 - 다중 제약

$$y_t = \beta_1 + \beta_2 X_{t2} + \beta_3 X_{t3} + \beta_4 X_{t4} + \varepsilon_t$$

$H_0: \beta_2 = 0, \beta_4 = 0$
 $H_1: H_0 \text{ not true}$

$$F = \frac{(SSE_R - SSE_U)/J}{SSE_U/(T-K)}$$

First run the restricted regression by dropping X_{t2} and X_{t4} to get SSE_R .
 Next run unrestricted regression to get SSE_U .

$df_n = J = 2$
 $df_d = T - K$

F-검정 계량경제학 8.14

실례 - 하나의 제약

$$y_t = \beta_1 + \beta_2 X_{t2} + \beta_3 X_{t3} + \varepsilon_t$$

$$F = \frac{(SSE_R - SSE_U)/J}{SSE_U/(T-K)}$$

$$= \frac{(1964.758 - 1805.168)/1}{1805.168/(52 - 3)}$$

$$= 4.33$$

$H_0: \beta_2 = 0$
 $H_1: \beta_2 \neq 0$

$df_n = J = 1$
 $df_d = T - K = 49$

By definition this is the t-statistic squared:
 $t = -2.081 \rightarrow F = t^2 = 4.33$

F-검정 계량경제학 8.17

실례 - 다중 제약

$$y_t = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} + \dots + \beta_K x_{tK} + \varepsilon_t$$

$H_0: \sum c_{1i} \beta_i = c_1, \sum c_{2i} \beta_i = c_2, \sum c_{3i} \beta_i = c_3$
 $H_1: H_0 \text{ not true}$

$$F = \frac{(SSE_R - SSE_U)/J}{SSE_U/(T-K)}$$

$df_n = J = 3$
 $df_d = T - K$

F-검정 계량경제학 8.15

실례 - 하나의 제약

$$y_t = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} + \dots + \beta_K x_{tK} + \varepsilon_t$$

$$F = \frac{(SSE_R - SSE_U)/J}{SSE_U/(T-K)}$$

$H_0: \sum c_i \beta_i = c_0$
 $H_1: H_0 \text{ not true}$

- SSE_R : the sum of squared errors when the restriction $\sum c_i \beta_i = c_0$ is imposed on the original model
- SSE_U : the sum of squared errors from the original model

$df_n = J = 1$
 $df_d = T - K$

F-검정 계량경제학 8.18

실례 - 모형의 유의성 검정

$$y_t = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} + \dots + \beta_K x_{tK} + \varepsilon_t$$

$H_0: \beta_2 = \beta_3 = \dots = \beta_K = 0$
 $H_1: H_0 \text{ not true}$

$$F = \frac{(SSE_R - SSE_U)/(K-1)}{SSE_U/(T-K)}$$

F-검정 계량경제학 8.19

실례 - 모형의 유의성 검정

	coefficient	std. error	t-ratio	p-value
const	104.786	6.48272	16.16	2.84e-021 ***
p	-6.64193	3.19119	-2.091	0.0427 **
a	2.98430	0.166936	17.88	4.11e-023 ***

Mean dependent var 120.3231 S.D. dependent var 16.31873
 Sum squared resid 1805.168 S.E. of regression 6.069611
 R-squared 0.967095 Adjusted R-squared 0.961660
 F(2, 49) 159.8280 F-value(F) 3.37e+22
 Log-likelihood -166.0111 Akaike criterion 338.0222
 Schwarz criterion 343.8759 Hannan-Quinn 340.2664

비표본정보의 이용 계량경제학 8.22

어떤 생산과정이 규모에 대한 수익불변인 Cobb-Douglas로 가정된다고 하자

$$\ln(y_t) = \beta_1 + \beta_2 \ln(X_{t2}) + \beta_3 \ln(X_{t3}) + \beta_4 \ln(X_{t4}) + \varepsilon_t$$

단 $\beta_2 + \beta_3 + \beta_4 = 1 \rightarrow \beta_4 = (1 - \beta_2 - \beta_3)$

$$\ln(y_t/X_{t4}) = \beta_1 + \beta_2 \ln(X_{t2}/X_{t4}) + \beta_3 \ln(X_{t3}/X_{t4}) + \varepsilon_t$$

$$y_t^* = \beta_1 + \beta_2 X_{t2}^* + \beta_3 X_{t3}^* + \varepsilon_t$$

변형된 모형에 대해 최소제곱추정을 함

F-검정 계량경제학 8.20

실례 - 모형의 유의성 검정

	coefficient	std. error	t-ratio	p-value
const	120.323	2.26300	53.17	2.43e-046 ***

Mean dependent var 120.3231 S.D. dependent var 16.31873
 Sum squared resid 13581.35 S.E. of regression 16.31873
 R-squared 0.000000 Adjusted R-squared 0.000000
 Log-likelihood -218.4802 Akaike criterion 438.9605
 Schwarz criterion 440.9117 Hannan-Quinn 439.7085

$$\frac{(13581.35 - 1805.168)/2}{1805.168/(52 - 3)} = 159.828$$

비표본정보의 이용 계량경제학 8.23

제한최소제곱추정

- 제한최소제곱추정량(restricted least squares estimator)은 부과한 제약이 정확히 사실이 아니면 불편추정량이 되지 못함
- 제한최소제곱추정량은 그 제약이 옳든 옳지 않든간에 그 분산이 원래의 모형에 대한 최소제곱추정량에 비해 작음
- 자료에 대한 추가적인 정보를 이용하고자 할 때 대개 불편성을 희생하는 대가로 분산을 줄여주게 됨

F-검정 v. χ^2 검정 계량경제학 8.21

χ^2 검정

$$V_1 = \frac{SSE_R - SSE_U}{\sigma^2} \sim \chi^2_{(J)}$$

여기서 σ^2 을 그 추정량($\hat{\sigma}^2$)으로 대체하면 V_1 도 같은 가설에 대한 검정통계량으로 사용할 수 있음

이 검정통계량은 근사적으로 χ^2 분포를 함 (χ^2 검정)

$$\rightarrow \frac{SSE_R - SSE_U}{\hat{\sigma}^2} = \chi^2_{(J)}, \text{ under the null}$$

소표본 분포를 이용한 F검정이 바람직하나, 정규분포의 가정을 버린다면 둘 다 근사적 분포를 이용

이 경우에도 F검정의 소표본 성질이 나타나, 관측치가 커짐에 따라 F검정은 χ^2 검정으로 수렴함을 보일 수 있음

원점을 통과하는 회귀식 계량경제학 8.24

상수항이 없거나 0인 모형.
 i.e., $y_t = \beta_2 x_t + \varepsilon_t$

LS 추정

$$b_2 = \frac{\sum x_t y_t}{\sum x_t^2} \quad \text{and} \quad \text{Var}(b_2) = \frac{\sigma^2}{\sum x_t^2}$$

and $\hat{\sigma}^2 = \frac{\sum \hat{\varepsilon}_t^2}{T-1}$

원점을 통과하는 회귀식의 특징

계량경제학 8.25

- $\sum \hat{\epsilon}_t$ 반드시 0일 이유는 없음
- R^2 모형의 통계적 성질을 나타내는 척도로 부적절하게 됨(계산 routine에 따라 음의 값이 나타날 수도 있음)
- $SST \neq SSR + SSE$
- df 상수항을 포함하지 않음, i.e., (T-K+1)

In practice:

- 매우 강한 선형적 혹은 이론적인 근거가 뒷받침 될 경우가 아니라면 원래의 상수항이 존재하는 모형을 사용
- 회귀모형에 상수항을 포함하고 그것이 통계적으로 유의하지 않는 경우 그것을 제거하고 다시 회귀분석을 할 수도 있음

모형의 설정

계량경제학 8.28

모형 설정의 핵심

- 모형에 어떠한 설명변수들을 포함 시킬 것인가?
- 포함되는 설명변수들과 종속변수는 어떠한 함수형태로 그 관계를 표현할 것인가?
- 모형의 통계적 가정들이 적절하다고 볼 수 있는가?
 - 다중회귀모형에 국한해서 본다면, 1-5의 가정이 충족된다고 볼 수 있는가?

모형의 설정

계량경제학 8.26

- 모형을 선택할 때 고려해야 하는 중요한 점은 무엇인가?
- 잘못된 모형의 선택으로 인한 결과는 무엇인가?
- 모형이 적절한지 여부를 평가할 방법이 있는가?

모형의 설정

계량경제학 8.29

누락된 변수와 관련없는 변수

True Model: $y_t = \beta_1 + \beta_2 X_{t2} + \beta_3 X_{t3} + \epsilon_t$

Estimate: $y_t = \beta_1 + \beta_2 X_{t2} + v_t$

- 옳지 않는 $\beta_3=0$ 이라는 제약을 모형에 부과한 것
- 적절한 변수의 “누락(Omission)”은 모형에 대한 잘못된 제약의 부과와 마찬가지로
- 불편추정량을 얻을 수 없으며 분산의 크기는 감소하게 됨
- 통제변수(control var.): 예컨대, X_2 가 분석 대상 변수일 때, 변수 누락으로 인한 편의를 방지하기 위해 포함되는 변수들

모형의 설정

계량경제학 8.27

인과관계 분석 모형 vs. 예측 모형

- 인과관계 분석: 설명변수의 변화가 종속변수의 평균(체계)적 변화에 미치는 영향의 크기를 알고자 함 (ex 정책 효과 분석)
 - 설명변수에서의 한 단위 변화가 다른 요인들이 일정할 때 종속변수의 평균에 미치는 영향을 고집여 내어야 함 (외생성이 중요)
- 예측 모형: 주어진 정보를 토대로 종속변수 값을 예측하고자 함
 - 이 경우 종속변수와 높은 상관성이 있는 변수들의 선택이 중요 (외생성 여부보다는 예측력을 높여주는 것이 중요)

모형의 설정

계량경제학 8.30

누락된 변수와 관련없는 변수

Estimate: $y_t = \beta_1 + \beta_2 X_{t2} + v_t \Rightarrow b_2^*$

Estimate: $y_t = \beta_1 + \beta_2 X_{t2} + \beta_3 X_{t3} + \epsilon_t \Rightarrow b_2$

$$E(b_2^*) = \beta_2 + \beta_3 \frac{\sum (x_{t2} - \bar{x}_2)(x_{t3} - \bar{x}_3) / (T-1)}{\sum (x_{t2} - \bar{x}_2)^2 / (T-1)}$$

$$= \beta_2 + \beta_3 \frac{\text{cov}(x_{t2}, x_{t3})}{\text{var}(x_{t2})} \neq \beta_2, \text{ Try to show this!}$$

$$\text{Var}(b_2^*) = \frac{\sigma_v^2}{\sum (x_{t2} - \bar{x}_2)^2}$$

$$\leq \frac{\sigma_\epsilon^2}{(1-r_{23}^2) \sum (x_{t2} - \bar{x}_2)^2} = \text{Var}(b_2)$$

$$\Rightarrow \text{Var}(v_t) = \text{Var}(\beta_3 x_{t3} + \epsilon_t) = \text{Var}(\epsilon_t)$$

- b_2^* 가 여전히 불편 추정량 인 경우:
 - X_{t2} 와 X_{t3} 가 상관되어 있지 않음
 - β_3 가 0일 때

누락된 변수와 관련없는 변수

True Model: $y_t = \beta_1 + \beta_2 X_{t2} + \varepsilon_t$

Estimate: $y_t = \beta_1 + \beta_2 X_{t2} + \beta_3 X_{t3} + \omega_t$

- In fact, $\beta_3=0$.
- 적절하지 않은 변수의 “포함”은 최소제곱 추정량의 불편성에 영향을 주지 않음
- 하지만 올바른 모형을 통해 얻는 추정량에 비해 b_1 와 b_2 의 분산이 더 커지게 됨

공선성

‘독립변수’는 설명변수가 오차항과 독립임을 의미하는 것이며 다른 설명변수들과 독립임을 의미하는 것은 아님

자료가 창출된 암묵적 실험에 대해 경제학자들은 대개 아무런 통제를 할 수 없기 때문에 설명변수들이 같이 움직임으로 인해 개별 설명변수들의 영향을 분리해내는 것을 문제가 되게 하는 경우가 종종 있음

- RESET(Regression Equation Specification Error Test)은 모형 설정에 있어서 변수의 누락과 잘못된 함수형태를 탐지하기 위해 고안됨
- 다음과 같이 설정된 추정식이 주어졌다면,

$$y_t = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} + e_t$$

- 최소제곱추정을 통해 다음의 적합값(fitted value)를 얻을 수 있음

$$\hat{y}_t = b_1 + b_2 x_{t2} + b_3 x_{t3}$$

- 이를 이용하여 다음과 같은 인위적인 회귀식들을 추정할 수 있음

$$y_t = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} + \gamma_1 \hat{y}_t^2 + e_t \quad (1)$$

$$y_t = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} + \gamma_1 \hat{y}_t^2 + \gamma_2 \hat{y}_t^3 + e_t \quad (2)$$

공선성의 효과

높은 공선성은 다음과 같은 문제를 일으킴

1. 공선성이 정확(exact)한 경우 최소제곱추정 결과를 얻을 수 없음 – 기본 가정 5
2. 표준오차가 커지고 따라서 신뢰구간이 넓어짐
3. 결정계수의 값이 높고 모형의 유의성에 대한 F값이 높음에도 불구하고 t값들이 유의성 없게 나타남
4. 추정치(그리고 그 유의성)들이 관측치 혹은 유의성 없는 변수들의 삭제 혹은 추가에 민감하게 변화함
5. 두 설명변수가 (공선성에서 암시되는) 일정 비율로 주어지는 경우 종속변수에 대한 예측력은 괜찮으나 다른 비율로 주어지는 경우 예측력이 나쁨

- (1)에서 모형설정오류에 대한 검정은 다음의 가설 검정으로 이루어짐
 $H_0: \gamma_1 = 0 \quad H_1: \gamma_1 \neq 0$
- (2)에서 모형설정오류에 대한 검정은 다음의 가설 검정으로 이루어짐
 $H_0: \gamma_1 = \gamma_2 = 0 \quad H_1: o.w.$
- 여기서 귀무가설의 기각은 원래 모형이 적절하지 못하며 개선의 여지가 있음을 의미함. 반면에 귀무가설을 기각하지 못하면 이는 이 검정을 통해 모형설정의 오류를 탐지할 수 없었음을 의미함
- 포함되는 적합값의 차수는 예시로 든 1차, 2차 보다 더 높아질 수도 있음
- Intuition? :

공선성의 확인

높은 공선성의 증거는 다음과 같음:

1. 두 설명변수간에 높은 상관계수 (> 0.8 or 0.9)
2. 하나의 설명변수를 다른 설명변수들에 대해 회귀 분석(보조적 회귀, auxiliary regression)을 차례로 수행할 때 나타나는 높은 결정계수의 값 (> 0.8)
3. t값들은 유의하지 못하는데, F값은 통계적으로 유의할 경우

공선성의 완화

높은 공선성은 표준적 가정에 대한 위반은 아니며 그 보다는 표본에 설명변수의 개별적 영향에 대한 정보가 부족함을 나타냄:

1. 더 나은 정보를 가진 자료의 추가적 수집
2. 적절한 경제적 제약을 부과.
3. 정당성이 확보되는 경우 통계적 제약을 부과.
4. 이 모든게 실패할 경우, 모형의 문제점이 이러한 공선성의 문제로 인한 것 (혹은 그로 인한 것이 아니라는 것)을 지적함

예측

$$y_t = \beta_1 + \beta_2 X_{t2} + \beta_3 X_{t3} + \varepsilon_t$$

Given a set of values for the explanatory variables, (X_2 X_3), the best linear unbiased predictor of y is given by:

$$\hat{y}_0 = b_1 + b_2 X_{02} + b_3 X_{03}$$

This predictor is **unbiased** in the sense that the **expectation of the forecast error is zero.**

$$f = (y_0 - \hat{y}_0) \Rightarrow E(f) = 0$$

예측

- The predictor is best in that the variance of the forecast error of any other linear and unbiased predictor of y_0 , is larger than $\text{var}(f) = \text{var}(y_0 - \hat{y}_0)$

$$\bullet \frac{f}{\text{se}(f)} = \frac{y_0 - \hat{y}_0}{\sqrt{\hat{\text{var}}(y_0 - \hat{y}_0)}} \sim t_{(T-K)}$$

- A $100(1-\alpha)\%$ interval predictor for y_0 is

$$\hat{y}_0 \pm t_c \text{se}(f)$$

, where t_c is a critical value from the $t_{(T-K)}$ distribution.