*Principles of Econometrics (3e)*

# Ch. 6 다중회귀모형에 관한 추가적인 논의

2013년 1학기

윤성민

- **6장의 주요 내용**

- 다중회귀모형의 모수에 관한 둘 이상의 가설로 구성된 귀무가설을 동시에 검정하는 경우 (**결합가설의 검정**)

⇒ F-검정

- 표본의 정보 이외에 **비표본 정보**도 함께 이용하는 경우

⇒ 제한 최소제곱법

- 모형 설정의 오류를 찾는 방법

⇒ **RESET 검정**

- **다중공선성** 문제의 탐지와 해결 방법

## 6.1 F-검정

<햄버거 체인점 사례>

- **제한되지 않은 모형(U)** $S_i = \beta_1 + \beta_2 P_i + \beta_3 A_i + e_i$

- 가설이 $H_0 : \beta_2 = 0 \;\; H_1 : \beta_2 \neq 0$ 로 설정된다면 $\Rightarrow$ t-검정

- **제한된 모형(R)** $S_i = \beta_1 + \beta_3 A_i + e_i \qquad (\beta_2 = 0)$

- 귀무가설이 참이라면 $SSE_R = SSE_U$

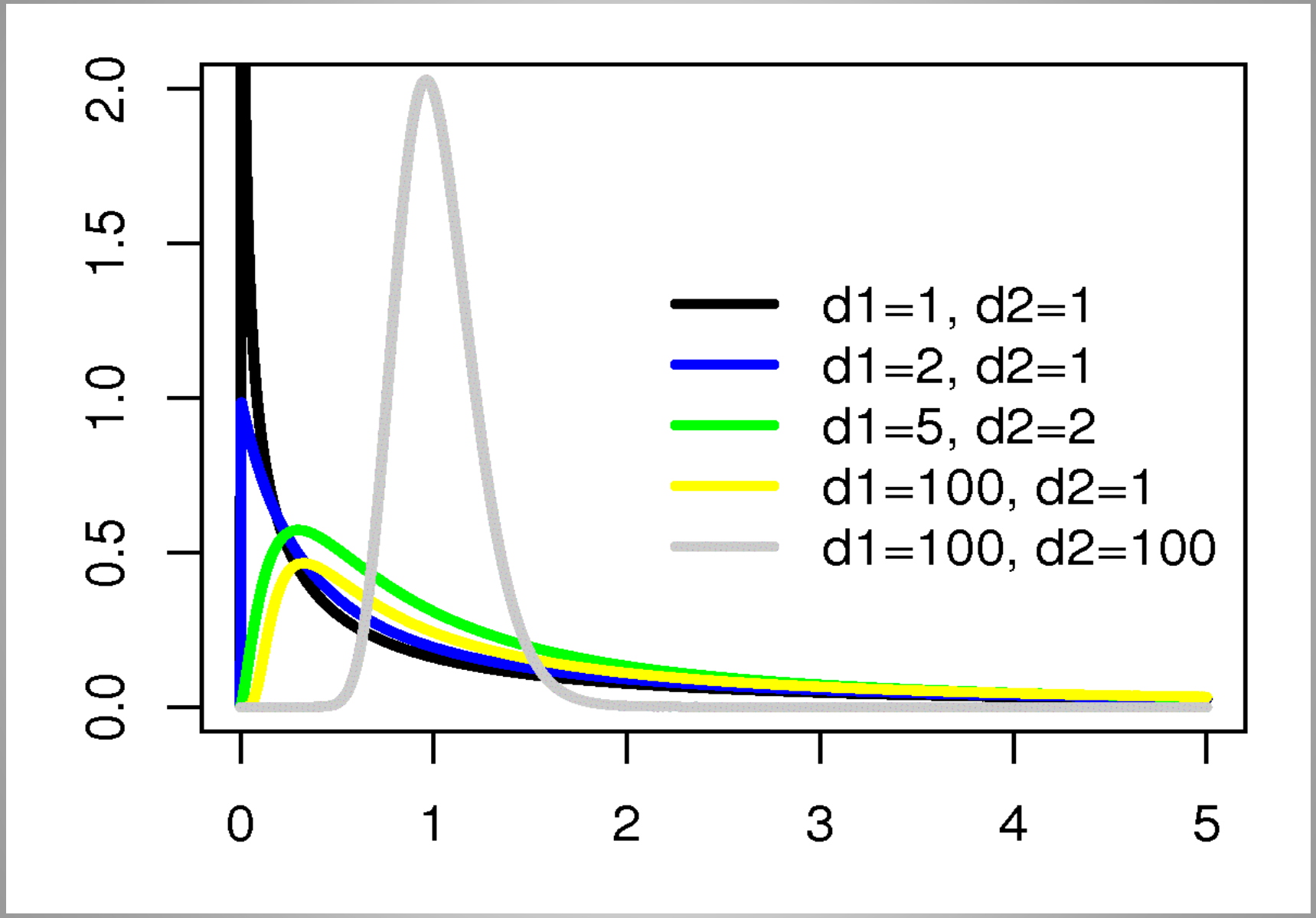
- 귀무가설이 거짓이라면 $SSE_R > SSE_U$

## ▪ F-분포

- F-distribution

  : 각각 자유도가 $m_1, m_2$ 인 카이제곱분포를 갖는 두 확률변수 $(V_1, V_2)$ 의 비율은 F-분포를 따름

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


0.0
0.5
1.0
1.5
2.0
0
1
2
3
4
5
d1=1, d2=1
d1=2, d2=1
d1=5, d2=2
d1=100, d2=1
d1=100, d2=100

## ▪ F-통계량

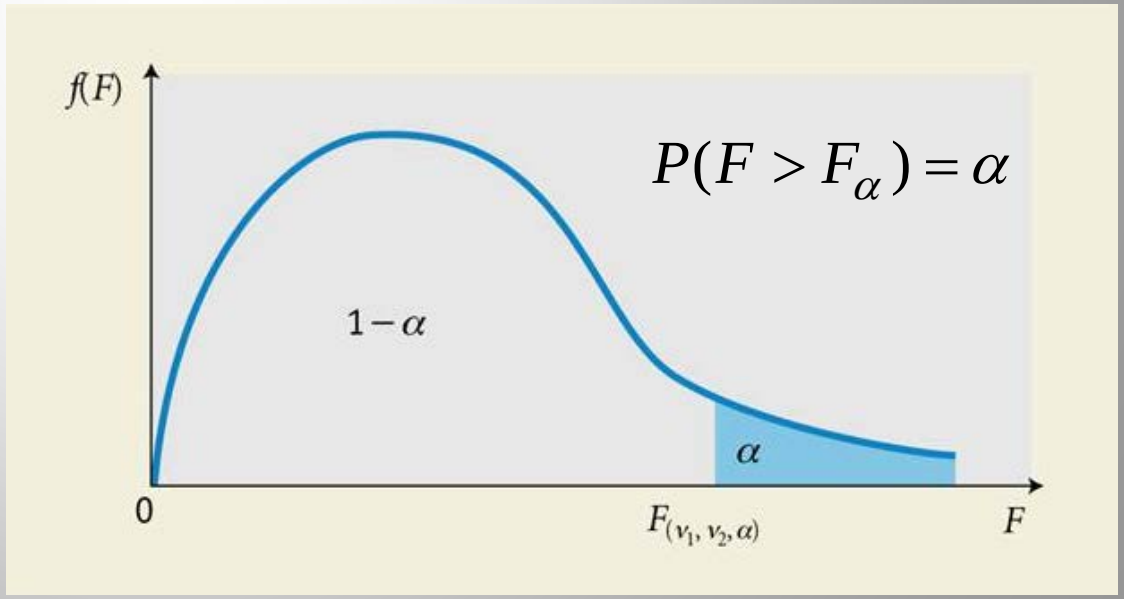
$$SSE = \sum \hat{e}_t^2 \quad \sim \quad \chi^2(N-K)$$

$$F = \frac{(SSE_R - SSE_U)/J}{SSE_U/(N-K)} \quad \sim \quad F(J, N-K)$$

- 모형에 설정된 제한이 부적절할수록 $(SSE_R - SSE_U)$ 커짐

  ⇒ F-값 커짐

  ⇒ 귀무가설 기각



$\alpha=0.05$가 되는 임계값은 $F_{(20,\ 10,\ 0.05)}=2.77$이다.

$\alpha = 0.05$

| 분모 자유도 $\nu_2$ | 분자 자유도 $\nu_1$ | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 12 | 15 | 20 | 24 | 30 | 40 | 60 | 120 | ∞ |
| 1 | 161.4 | 199.5 | 215.7 | 224.6 | 230.2 | 234.0 | 236.8 | 238.9 | 240.5 | 241.9 | 243.9 | 245.9 | 248.0 | 249.1 | 250.1 | 251.1 | 252.2 | 253.3 | 254.3 |
| 2 | 18.51 | 19.00 | 19.16 | 19.25 | 19.30 | 19.33 | 19.35 | 19.37 | 19.38 | 19.40 | 19.41 | 19.43 | 19.45 | 19.45 | 19.46 | 19.47 | 19.48 | 19.49 | 19.50 |
| 3 | 10.13 | 9.55 | 9.28 | 9.12 | 9.01 | 8.94 | 8.89 | 8.85 | 8.81 | 8.79 | 8.74 | 8.70 | 8.66 | 8.64 | 8.62 | 8.59 | 8.57 | 8.55 | 8.53 |
| 4 | 7.71 | 6.94 | 6.59 | 6.39 | 6.26 | 6.16 | 6.09 | 6.04 | 6.00 | 5.96 | 5.91 | 5.86 | 5.80 | 5.77 | 5.75 | 5.72 | 5.69 | 5.66 | 5.63 |
| 5 | 6.61 | 5.79 | 5.41 | 5.19 | 5.05 | 4.95 | 4.88 | 4.82 | 4.77 | 4.74 | 4.68 | 4.62 | 4.56 | 4.53 | 4.50 | 4.46 | 4.43 | 4.40 | 4.36 |
| 6 | 5.99 | 5.14 | 4.76 | 4.53 | 4.39 | 4.28 | 4.21 | 4.15 | 4.10 | 4.06 | 4.00 | 3.94 | 3.87 | 3.84 | 3.81 | 3.77 | 3.74 | 3.70 | 3.67 |
| 7 | 5.59 | 4.74 | 4.35 | 4.12 | 3.97 | 3.87 | 3.79 | 3.73 | 3.68 | 3.64 | 3.57 | 3.51 | 3.44 | 3.41 | 3.38 | 3.34 | 3.30 | 3.27 | 3.23 |
| 8 | 5.32 | 4.46 | 4.07 | 3.84 | 3.69 | 3.58 | 3.50 | 3.44 | 3.39 | 3.35 | 3.28 | 3.22 | 3.15 | 3.12 | 3.08 | 3.04 | 3.01 | 2.97 | 2.93 |
| 9 | 5.12 | 4.26 | 3.86 | 3.63 | 3.48 | 3.37 | 3.29 | 3.23 | 3.18 | 3.14 | 3.07 | 3.01 | 2.94 | 2.90 | 2.86 | 2.83 | 2.79 | 2.75 | 2.71 |
| 10 | 4.96 | 4.10 | 3.71 | 3.48 | 3.33 | 3.22 | 3.14 | 3.07 | 3.02 | 2.98 | 2.91 | 2.85 | 2.77 | 2.74 | 2.70 | 2.66 | 2.62 | 2.58 | 2.54 |
| 11 | 4.84 | 3.98 | 3.59 | 3.36 | 3.20 | 3.09 | 3.01 | 2.95 | 2.90 | 2.85 | 2.79 | 2.72 | 2.65 | 2.61 | 2.57 | 2.53 | 2.49 | 2.45 | 2.40 |
| 12 | 4.75 | 3.89 | 3.49 | 3.26 | 3.11 | 3.00 | 2.91 | 2.85 | 2.80 | 2.75 | 2.69 | 2.62 | 2.54 | 2.51 | 2.47 | 2.43 | 2.38 | 2.34 | 2.30 |
| 13 | 4.67 | 3.81 | 3.41 | 3.18 | 3.03 | 2.92 | 2.83 | 2.77 | 2.71 | 2.67 | 2.60 | 2.53 | 2.46 | 2.42 | 2.38 | 2.34 | 2.30 | 2.25 | 2.21 |
| 14 | 4.60 | 3.74 | 3.34 | 3.11 | 2.96 | 2.85 | 2.76 | 2.70 | 2.65 | 2.60 | 2.53 | 2.46 | 2.39 | 2.35 | 2.31 | 2.27 | 2.22 | 2.18 | 2.13 |

## ▪ single restriction F-검정 (단일가설 검정 경우)

- 가설: $H_0 : \beta_2 = 0 \quad H_1 : \beta_2 \neq 0$ (총수입은 가격에 의존하는가?)
- 검정통계량: $F = \dfrac{(SSE_R - SSE_U)/J}{SSE_U/(N-K)} \sim F(J, N-K)$
- 유의수준 선택: $\alpha = 0.05$
- 자유도 및 임계값: 단일가설이므로 J=1, (N-K)=75-3=72이므로

  임계값은 $F_c = F_{(1,72;0.05)} = 3.97$ =FINV(0.05,1,72)
- F-통계량: $F = \dfrac{(SSE_R - SSE_U)/J}{SSE_U/(N-K)} = \dfrac{(2961.827 - 1718.943)/1}{1718.943/(75-3)} = 52.06$
- 검정결과: $F = 52.06 > 3.97 = F_c$ 따라서 귀무가설 기각(가격에 의존)
- P-값 계산: $p = P[F_{(1,72)} \geq 52.06] = .0000$ =FDIST(52.06,1,72)

$$p = 0.0000 < 0.05 = \alpha$$

- 햄버거 체인점의 총수입은 가격에 의존하는가? **(t-검정)**

- 가설: $H_0 : \beta_2 = 0 \quad H_1 : \beta_2 \neq 0$
- 검정통계량: $t = \dfrac{b_2}{se(b_2)} \sim t_{(N-K)}$
- 유의수준 선택: $\alpha = 0.05$ =TINV(0.05,72)
- 임계값 및 기각역: 자유도=75-3=72이므로 임계값은 $t_c = \pm 1.993$

  기각역은 $|t| \geq 1.993$
- t-통계량 및 P-값 계산: $t = \dfrac{b_2}{se(b_2)} = \dfrac{-7.908}{1.096} = -7.215$

  $P\left(t_{(72)} > 7.215\right) + P\left(t_{(72)} < -7.215\right) = 2 \times (2.2 \times 10^{-10}) = 0.000$
- 검정결과: $-7.215 < -1.993$, $p = 0.000 < 0.05 = \alpha$ =TDIST(−7.215,72,2)

따라서 귀무가설 기각, 즉 수입은 가격에 의존함

$F = 52.06 = t^2 = (-7.215)^2$ $\quad F_c = 3.97 = t_c^2 = (1.993)^2$

## 6.2 모형의 유의성 검정

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

- **단일 추정치의 유의성 검정 ⇒ t-검정** (F-검정과 동일한 결과)

  $H_0$ : 어느 한 $\beta_k = 0$

  $H_1 : \beta_k \neq 0$

- **회귀모형의 전반적인 유의성 검정**

  **= 결합가설의 검정 ⇒ F-검정**

  $H_0 : \beta_2 = 0, \beta_3 = 0, \ldots, \beta_K = 0$

  $H_1$ : *at least one* of the $\beta_k$ is nonzero

- **multiple restriction F-검정 (결합가설 검정 경우)**

$$y_i = \beta_1 + x_{i2}\beta_2 + x_{i3}\beta_3 + \cdots + x_{iK}\beta_K + e_i$$

- (귀무가설) $H_0 : \beta_2 = 0 \;\; and \;\; \beta_3 = 0 \;\; \cdots \;\; and \;\; \beta_K = 0$
- (대립가설) $H_1 : \beta_k$ 들 중 적어도 하나는 0이 아니다
- 이 경우 제한된 모형: $y_i = \beta_1 + e_i$
- 이 모형의 OLS 추정량 $b_1^* = \sum Y_t / T = \bar{Y}$

$$SSE_R = \sum_{i=1}^{N} (y_i - b_1^*)^2 = \sum_{i=1}^{N} (y_i - \bar{y})^2 = SST$$

$$SSE_U = SSE \qquad J = K - 1$$

$$F = \frac{(SSE_R - SSE_U)/J}{SSE_U/(N-K)} = \frac{(SST - SSE)/(K-1)}{SSE/(N-K)} \quad \sim \quad F(K-1, N-K)$$

## ▪ 햄버거 체인점, 결합가설 검정 (회귀식의 전반적 유의성 검정)

$$S_i = \beta_1 + \beta_2 P_i + \beta_3 A_i + e_i$$

- 귀무가설: $H_0 : \beta_2 = 0 \;\; and \;\; \beta_3 = 0$
- 대립가설: $H_1 : \beta_2 \neq 0 \;\; or \;\; \beta_3 \neq 0$

분산 분석

| | 자유도 | 제곱합 | 제곱 평균 | F 비 | 유의한 F |
|---|---|---|---|---|---|
| 회귀 | 2 | 1396.539 | 698.2695 | 29.24786 | 5.04E-10 |
| 잔차 | 72 | 1718.943 | 23.87421 | | |
| 계 | 74 | 3115.482 | | | |

$$F = \frac{(SST - SSE)/(K-1)}{SSE/(N-K)} = \frac{(3115.482 - 1718.943)/2}{1718.943/(75-3)} = 29.25$$

- $F_c = F(2, 72; 0.05) = 3.12$ $\qquad 29.25 > 3.12$ 귀무가설 기각함

$$p = P[F_{(2,72)} \geq 29.25] = .0000$$

# ▪ Eviews output

Dependent Variable: SALES
Method: Least Squares
Date: 11/04/10 Time: 23:29
Sample: 1 75
Included observations: 75
SALES=C(1)+C(2)*PRICE+C(3)*ADVERT

| | Coefficient | Std. Error | t-Statistic | Prob. |
|---|---|---|---|---|
| C(1) | 118.9136 | 6.351638 | 18.72172 | 0.0000 |
| C(2) | -7.907854 | 1.095993 | -7.215241 | 0.0000 |
| C(3) | 1.862584 | 0.683195 | 2.726283 | 0.0080 |

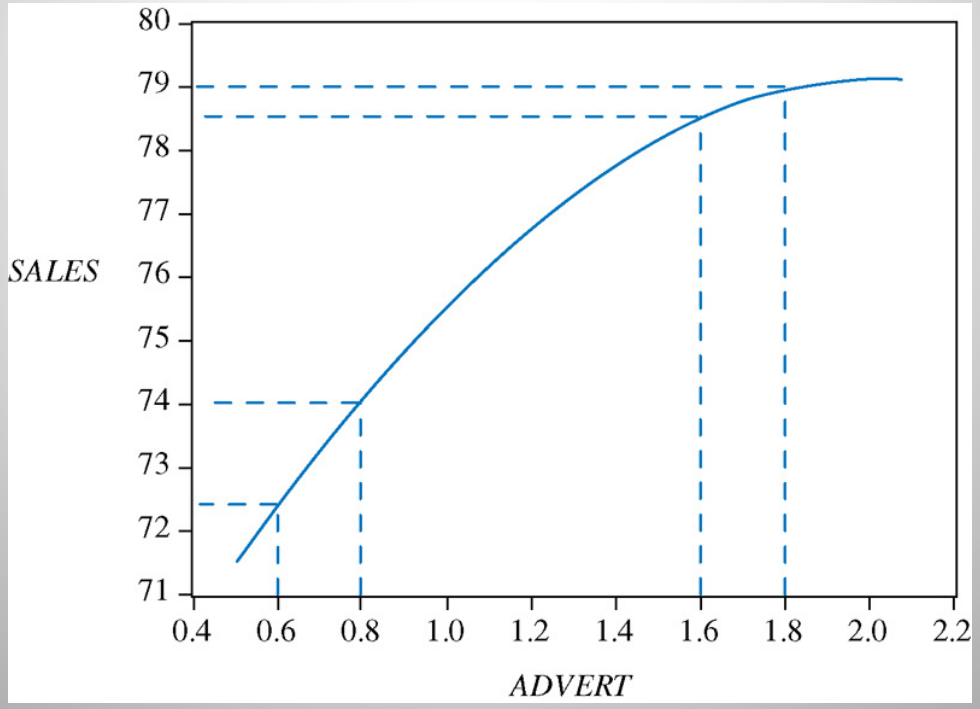
| | | | |
|---|---|---|---|
| R-squared | 0.448258 | Mean dependent var | 77.37467 |
| Adjusted R-squared | 0.432932 | S.D. dependent var | 6.488537 |
| S.E. of regression | 4.886124 | Akaike info criterion | 6.049854 |
| Sum squared resid | 1718.943 | Schwarz criterion | 6.142553 |
| Log likelihood | -223.8695 | Hannan-Quinn criter. | 6.086868 |
| F-statistic | 29.24786 | Durbin-Watson stat | 2.183037 |
| Prob(F-statistic) | 0.000000 | | |

## 6.3 확장된 모형

$$S = \beta_1 + \beta_2 P + \beta_3 A + e$$

✓ 광고비 지출을 늘리면 매출액이 계속 비례적으로 증가할 것인가?

- 현실적으로 광고비의 효과에는 수확체감의 법칙이 작용할 것 같음



## ▪ 확장된 모형

- 광고비 지출이 매출액 증대에 미치는 효과에 수확체감이 작용하는 경우, 다음과 같은 모형이 적합할 수 있음

$$S = \beta_1 + \beta_2 P + \beta_3 A + \beta_4 A^2 + e$$

$$\frac{\Delta E(S)}{\Delta A}_{(P \text{ held constant})} = \frac{\partial E(S)}{\partial A} = \beta_3 + 2\beta_4 A$$

- 이 경우 $\beta_3 > 0, \quad \beta_4 < 0$ 으로 예상됨

## ▪ 확장된 모형의 추정결과

- 추정방법: 아래와 같이 변수를 변환하여 OLS 적용함

$$S = \beta_1 + \beta_2 P + \beta_3 A + \beta_4 A^2 + e$$

$$y_i = S_i, \quad x_{i2} = P_i, \quad x_{i3} = A_i, \quad x_{i4} = A_i^2$$

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + e$$

- 추정결과

$$\hat{S}_i = 109.72 - 7.640 P_i + 12.151 A_i - 2.768 A_i^2$$

$$(\text{se}) \quad (6.80) \quad (1.046) \qquad (3.556) \quad (0.941)$$

## 6.4 경제적 가설의 검정

### 6.4.1 광고의 유용성

✓ 광고는 매출에 영향을 미치는가?

- 제한되지 않은 모형 $S_i = \beta_1 + \beta_2 P_i + \beta_3 A_i + \beta_4 A_i^2 + e_i \quad (U)$

$$H_0 : \beta_3 = 0, \beta_4 = 0$$

$$H_1 : \beta_3 \neq 0 \text{ or } \beta_4 \neq 0$$

- 귀무가설이 옳다면(즉, 광고가 매출에 아무런 도움이 안 된다면)

$$S_i = \beta_1 + \beta_2 P_i + e_i \quad (R)$$

- 햄버거 체인점의 광고는 총수입에 영향을 미치는가? **(F-검정)**

$$S_i = \beta_1 + \beta_2 P_i + \beta_3 A_i + \beta_4 A_i^2 + e_i$$

- 가설: $H_0 : \beta_3 = 0, \beta_4 = 0 \quad H_1 : \beta_3 \neq 0 \text{ or } \beta_4 \neq 0$
- 검정통계량: $F = \dfrac{(SSE_R - SSE_U)/2}{SSE_U/(75-4)} \sim F_{(2,71)}$
- 유의수준 선택: $\alpha = 0.05$
- 임계값 및 기각역: 자유도=75-4=71이므로 임계값은 $F_c = F_{(.95,2,71)} = 3.126$

  기각역은 $F > 3.126$

- F-통계량 및 P-값 계산: $F = \dfrac{(1896.391 - 1532.084)/2}{1532.084/(75-4)} = 8.44$

$$P[F_{(2,71)} > 8.44] = .0005$$

- 검정결과: $F_c = 3.126 < F = 8.44$

  따라서 귀무가설 기각, 즉 광고는 매출액에 영향 미침

## 6.4.2 광고의 최적 수준

$$S_i = \beta_1 + \beta_2 P_i + \beta_3 A_i + \beta_4 A_i^2 + e_i$$

- 경제이론에 따르면 한계수익이 한계비용보다 큰 모든 행위를 해야 함
- 광고를 한 단위 증가시킴에 따른 한계수익 = 기대 매출액의 증가

$$\frac{\Delta E(S)}{\Delta A}_{(P \text{ held constant})} = \beta_3 + 2\beta_4 A$$

- 광고를 한 단위 증가시킴에 따른 한계비용 = 광고비(1 단위금액)
  (매출 증가에 따른 재료비, 인건비, 관리비 등은 없다고 가정함)
- 한계수입=한계비용은 아래의 조건이 충족될 때 나타남

$$\beta_3 + 2\beta_4 A_o = 1$$

- 추정치 대입, $(12.1512) + 2 \times (-2.76796)\hat{A}_o = 1 \Rightarrow \hat{A}_o = 2.104$
- 따라서 광고의 최적 수준은 $2,104 임

## 6.5 비표본 정보의 이용

- 비표본 정보(Nonsample Information)
  - 표본자료에 포함된 정보 이외에 추가적인 정보
  - 경제원칙, 경제이론, 경험 등에서 얻을 수 있음

➢ 모수를 추정할 때 정확성을 향상시킬 수 있음

## ▪ 비표본 정보의 이용: 맥주수요함수 예

$$q = f(p_B, p_L, p_R, m)$$

- $q$ : 맥주수요량, $P_B$ : 맥주가격, $P_L$ : 다른 술 가격
  $P_R$ : 기타 다른 모든 재화 가격, $m$ : 소득

$$\ln q = \beta_1 + \beta_2 \ln p_B + \beta_3 \ln p_L + \beta_4 \ln p_R + \beta_5 \ln m$$

- 화폐착각(money illusion)이 없다면, 모든 가격과 소득이 동일한 비율로 증가할 경우 수요량은 변화하지 않을 것임

$$\begin{aligned}\ln q &= \beta_1 + \beta_2 \ln(\lambda p_B) + \beta_3 \ln(\lambda p_L) + \beta_4 \ln(\lambda p_R) + \beta_5 \ln(\lambda m) \\ &= \beta_1 + \beta_2 \ln p_B + \beta_3 \ln p_L + \beta_4 \ln p_R + \beta_5 \ln m + (\beta_2 + \beta_3 + \beta_4 + \beta_5)\ln\lambda\end{aligned}$$

- 화폐착각이 없다면(소비자가 합리적이라면) $\beta_2 + \beta_3 + \beta_4 + \beta_5 = 0$

- **맥주 수요함수 추정에 사용된 통계자료의 기통통계값**

*Table 6.1* Summary Statistics for Data Used to Estimate Beer Demand

| | *Q* | *PB* | *PL* | *PR* | *I* |
|---|---|---|---|---|---|
| Sample mean | 56.11 | 3.08 | 8.37 | 1.25 | 32,602 |
| Median | 54.90 | 3.11 | 8.39 | 1.18 | 32,457 |
| Maximum | 81.70 | 4.07 | 9.52 | 1.73 | 41,593 |
| Minimum | 44.30 | 1.78 | 6.95 | 0.67 | 25,088 |
| Std. Dev. | 7.8574 | 0.6422 | 0.7696 | 0.2983 | 4,542 |

$q$ : 맥주수요량, $P_B$ : 맥주가격, $P_L$ : 다른 술 가격
$P_R$ : 기타 다른 모든 재화 가격, $m$ : 소득

$$\ln q = \beta_1 + \beta_2 \ln p_B + \beta_3 \ln p_L + \beta_4 \ln p_R + \beta_5 \ln m$$

$$\beta_2 + \beta_3 + \beta_4 + \beta_5 = 0$$

- **비표본 정보**인 다음 관계식을 원래 식에 대입

$$\beta_4 = -\beta_2 - \beta_3 - \beta_5$$

$$\begin{aligned}
\ln q_t &= \beta_1 + \beta_2 \ln p_{Bt} + \beta_3 \ln p_{Lt} + \left(-\beta_2 - \beta_3 - \beta_5\right) \ln p_{Rt} + \beta_5 \ln m_t + e_t \\
&= \beta_1 + \beta_2 \left(\ln p_{Bt} - \ln p_{Rt}\right) + \beta_3 \left(\ln p_{Lt} - \ln p_{Rt}\right) + \beta_5 \left(\ln m_t - \ln p_{Rt}\right) + e_t \\
&= \beta_1 + \beta_2 \ln\left(\frac{p_{Bt}}{p_{Rt}}\right) + \beta_3 \ln\left(\frac{p_{Lt}}{p_{Rt}}\right) + \beta_5 \ln\left(\frac{m_t}{p_{Rt}}\right) + e_t
\end{aligned}$$

$$\begin{aligned}
\ln \hat{q}_t = \; & -4.798 - 1.2994 \ln\left(\frac{p_{Bt}}{p_{Rt}}\right) + 0.1868 \ln\left(\frac{p_{Lt}}{p_{Rt}}\right) + 0.9458 \ln\left(\frac{m_t}{p_{Rt}}\right) \\
& (3.714) \;\; (0.166) \qquad\qquad (0.284) \qquad\qquad (0.427)
\end{aligned}$$

$$\begin{aligned}
b_4^* &= -b_2^* - b_3^* - b_5^* \\
&= -(-1.2994) - 0.1868 - 0.9458 \\
&= 0.1668
\end{aligned}$$

# ▪ 맥주수요함수 추청결과: linear model

Dependent Variable: Q
Method: Least Squares
Date: 04/23/11 Time: 16:22
Sample: 1 30
Included observations: 30
Q=C(1)+C(2)*PB+C(3)*PL+C(4)*PR+C(5)*M

| | Coefficient | Std. Error | t-Statistic | Prob. |
|---|---|---|---|---|
| C(1) | 82.15871 | 17.96176 | 4.574090 | 0.0001 |
| C(2) | -23.74260 | 5.429409 | -4.372962 | 0.0002 |
| C(3) | -4.077410 | 3.890489 | -1.048046 | 0.3046 |
| C(4) | 12.92434 | 4.163896 | 3.103906 | 0.0047 |
| C(5) | 0.001995 | 0.000776 | 2.570612 | 0.0165 |

| | | | |
|---|---|---|---|
| R-squared | 0.822118 | Mean dependent var | 56.11333 |
| Adjusted R-squared | 0.793657 | S.D. dependent var | 7.857381 |
| S.E. of regression | 3.569219 | Akaike info criterion | 5.533582 |
| Sum squared resid | 318.4831 | Schwarz criterion | 5.767115 |
| Log likelihood | -78.00374 | Hannan-Quinn criter. | 5.608292 |
| F-statistic | 28.88559 | Durbin-Watson stat | 2.508801 |
| Prob(F-statistic) | 0.000000 | | |

# 맥주수요함수 추청결과: log-log model

Dependent Variable: LOG(Q)
Method: Least Squares
Date: 04/23/11 Time: 16:37
Sample: 1 30
Included observations: 30
LOG(Q)=C(1)+C(2)*LOG(PB)+C(3)*LOG(PL)+C(4)*LOG(PR)+C(5)*LOG(M)

| | Coefficient | Std. Error | t-Statistic | Prob. |
|---|---|---|---|---|
| C(1) | -3.243238 | 3.743000 | -0.866481 | 0.3945 |
| C(2) | -1.020419 | 0.239042 | -4.268787 | 0.0002 |
| C(3) | -0.582934 | 0.560150 | -1.040674 | 0.3080 |
| C(4) | 0.209545 | 0.079693 | 2.629415 | 0.0144 |
| C(5) | 0.922864 | 0.415514 | 2.221016 | 0.0356 |

| | | | |
|---|---|---|---|
| R-squared | 0.825389 | Mean dependent var | 4.018531 |
| Adjusted R-squared | 0.797451 | S.D. dependent var | 0.133258 |
| S.E. of regression | 0.059973 | Akaike info criterion | -2.638823 |
| Sum squared resid | 0.089920 | Schwarz criterion | -2.405290 |
| Log likelihood | 44.58235 | Hannan-Quinn criter. | -2.564114 |
| F-statistic | 29.54377 | Durbin-Watson stat | 2.630645 |
| Prob(F-statistic) | 0.000000 | | |

# 맥주수요함수 추청결과: 비표본 정보 이용(1)

Dependent Variable: LOG(Q)
Method: Least Squares
Date: 04/23/11 Time: 16:50
Sample: 1 30
Included observations: 30
LOG(Q)=C(1)+C(2)*LOG(PB/PR)+C(3)*LOG(PL/PR)+C(5)*LOG(M/PR)

| | Coefficient | Std. Error | t-Statistic | Prob. |
|---|---|---|---|---|
| C(1) | -4.797798 | 3.713905 | -1.291847 | 0.2078 |
| C(2) | -1.299386 | 0.165738 | -7.840022 | 0.0000 |
| C(3) | 0.186816 | 0.284383 | 0.656916 | 0.5170 |
| C(5) | 0.945829 | 0.427047 | 2.214813 | 0.0357 |

| | | | |
|---|---|---|---|
| R-squared | 0.807949 | Mean dependent var | 4.018531 |
| Adjusted R-squared | 0.785789 | S.D. dependent var | 0.133258 |
| S.E. of regression | 0.061676 | Akaike info criterion | -2.610291 |
| Sum squared resid | 0.098901 | Schwarz criterion | -2.423465 |
| Log likelihood | 43.15437 | Hannan-Quinn criter. | -2.550524 |
| F-statistic | 36.46021 | Durbin-Watson stat | 2.686998 |
| Prob(F-statistic) | 0.000000 | | |

# 맥주수요함수 추청결과: 비표본 정보 이용(2)

Dependent Variable: LOG(Q)
Method: Least Squares
Date: 04/23/11 Time: 17:03
Sample: 1 30
Included observations: 30
LOG(Q)=C(1)+C(3)*LOG(PL/PB)+C(4)*LOG(PR/PB)+C(5)*LOG(M/PB)

| | Coefficient | Std. Error | t-Statistic | Prob. |
|---|---|---|---|---|
| C(1) | -4.797798 | 3.713905 | -1.291847 | 0.2078 |
| C(3) | 0.186816 | 0.284383 | 0.656916 | 0.5170 |
| C(4) | 0.166742 | 0.077075 | 2.163369 | 0.0399 |
| C(5) | 0.945829 | 0.427047 | 2.214813 | 0.0357 |

| | | | |
|---|---|---|---|
| R-squared | 0.807949 | Mean dependent var | 4.018531 |
| Adjusted R-squared | 0.785789 | S.D. dependent var | 0.133258 |
| S.E. of regression | 0.061676 | Akaike info criterion | -2.610291 |
| Sum squared resid | 0.098901 | Schwarz criterion | -2.423465 |
| Log likelihood | 43.15437 | Hannan-Quinn criter. | -2.550524 |
| F-statistic | 36.46021 | Durbin-Watson stat | 2.686998 |
| Prob(F-statistic) | 0.000000 | | |

## 6.6 모형 설정

**<설정된 모형에 따라 달라지는 문제들>**

- 모형의 모수를 추정하는 최선의 방법은 무엇인가?
- 모형의 모수에 관한 가설을 어떻게 설정하여야 하는가?
- 모형의 모수에 관한 가설을 어떻게 검정할 수 있는가?
- 예측은 어떻게 해야 하는가?

✓ 모형을 선택할 경우 어떤 사항을 중요하게 고려하여야 하나?

✓ 선택한 모형이 적절한지를 평가하는 방법은 무엇인가?

**<모형을 선택할 경우 고려해야 할 사항>**

(1) 모형에 포함되어야 할 설명변수의 선정

(2) 함수형태의 선택

(3) 다중회귀모형의 기본가정인 MR1~MR6의 준수 여부

## 6.6.1 주요 변수의 누락이 유발하는 문제 (omitted-variable problem)

- 올바른 모형이 다음과 같다고 하자

$$W_t = \beta_1 + \beta_2 E_t + \beta_3 M_t + e_t$$

$W$ : 피고용인 임금, $E$ : 경험, $M$ : 열의(motivation)

- 부적절한 모형(주요 설명변수 $M$ 이 누락된 모형)

$$W_t = \beta_1 + \beta_2 E_t + v_t$$

- 부적절한 제약 $\beta_3 = 0$ 이 부과된 셈

⇒ 누락된 변수($M$)와 포함된 변수($E$)가 완전 독립이 아니라면, $\beta_1, \beta_2$ 의 OLS 추정량은 분산이 감소하나 불편추정량이 못됨

✓ 추정된 계수가 예상하지 못한 부호를 갖거나 비현실적인 값을 가지는 원인이 되기도 함 (증명: 6장 부록6B)

## ▪ 가계소득 추정 사례

- 맞벌이 가계가 많아, 가계총소득(*FAMINC*)은 남편의 교육수준(*HEDU*)과 아내의 교육수준(*WEDU*)에 크게 의존한다고 하자
- 아래 자료로 추정

**Table 6.2** Summary Statistics for Data Used for Family Income Example

| | *FAMINC* | *HEDU* | *WEDU* | *KL6* | $X_5$ | $X_6$ |
|---|---|---|---|---|---|---|
| Sample mean | 91213 | 12.61 | 12.65 | 0.14 | 12.57 | 25.13 |
| Median | 83013 | 12 | 12 | 0 | 12.60 | 24.91 |
| Maximum | 344146 | 17 | 17 | 2 | 20.82 | 37.68 |
| Minimum | 9072 | 4 | 5 | 0 | 2.26 | 9.37 |
| Std. Dev. | 44147 | 3.035 | 2.285 | 0.392 | 3.427 | 5.052 |
| Correlation matrix | | | | | | |
| *FAMINC* | 1.000 | | | | | |
| *HEDU* | 0.355 | 1.000 | | | | |
| *WEDU* | 0.362 | 0.594 | 1.000 | | | |
| *KL6* | −0.072 | 0.105 | 0.129 | 1.000 | | |
| $X_5$ | 0.290 | 0.836 | 0.518 | 0.149 | 1.000 | |
| $X_6$ | 0.351 | 0.821 | 0.799 | 0.160 | 0.900 | 1.000 |

- **가계소득 추정 결과**

$$\widehat{FAMINC}_i = -5534 + 3132\,HEDU_i + 4523\,WEDU_i$$

| | | | |
|---|---|---|---|
| (se) | (11230) | (803) | (1066) |
| ($p$-value) | (.622) | (.000) | (.000) |

- 남편이 교육을 1년 더 받을 경우 연간 가계소득은 $3,132 증가

- **중요 변수**($WEDU$)가 **누락된 경우**의 추정 결과

$$\widehat{FAMINC}_i = -26191 + 5155\,HEDU_i$$

| | | |
|---|---|---|
| (se) | (8541) | (658) |
| ($p$-value) | (.002) | (.000) |

- 남편이 교육을 1년 더 받을 경우 연간 가계소득은 $5,155 증가

  즉, 약 $2,000 과대평가 (불편추정량 아닌 것을 확인할 수 있음)

- **가계소득 추정 결과**

$$\widehat{FAMINC}_i = -5534 + 3132\,HEDU_i + 4523\,WEDU_i$$

| | | | |
|---|---|---|---|
| (se) | (11230) | (803) | (1066) |
| ($p$-value) | (.622) | (.000) | (.000) |

- 남편이 교육을 1년 더 받을 경우 연간 가계소득은 $3,132 증가

- **중요하지 않은 변수**(*KL6*, 6세 미만 자녀수)가 **누락된 경우**의 추정 결과

$$\widehat{FAMINC}_i = -7755 + 3211\,HEDU_i + 4777\,WEDU_i - 14311\,KL6_i$$

| | | | | |
|---|---|---|---|---|
| (se) | (11163) | (797) | (1061) | (5004) |
| ($p$-value) | (.488) | (.000) | (.000) | (.004) |

- 남편이 교육을 1년 더 받을 경우 연간 가계소득은 $3,211 증가

*HEDU*와 *WEDU* 추정치에 별 변화 없음 (불편추정량일 가능성 높음)

- *KL6*가 *HEDU* 및 *WEDU*와 독립이면, 누락되어도 전혀 영향 주지 않음

- **어떤 한 변수가 회귀모형에 포함되어야 하는지를 판단하는 방법**

- 기본원칙: 유의성 검정

추정치가 0이라는 귀무가설이 기각되면 포함시킴

<주의>

- 추정치가 0이라는 귀무가설이 기각되지 못하는 이유

(1) 해당 변수가 종속변수와 무관한 경우 ⇒ 제외시켜야 함

(2) 중요한 변수임에도 불구하고, 표본이 적절하지 못한 경우

⇒ 제외시키면, 나머지 추정치는 불편추정량이 되지 못함

✓ 가능한 많은 변수를 포함시키는 것이 좋은 전략인가?

Cf. stepwise 옵션

```
model  Y = X1 X2 X3 X4 / selection=stepwise ;
```

- **가계소득 추정 결과**

$$\widehat{FAMINC}_i = -5534 + 3132\,HEDU_i + 4523\,WEDU_i$$

| | | | |
|---|---|---|---|
| (se) | (11230) | (803) | (1066) |
| ($p$-value) | (.622) | (.000) | (.000) |

- 남편이 교육을 1년 더 받을 경우 연간 가계소득은 $3,132 증가

- **무관한 변수**(*KL6*, 6세 미만 자녀수)가 **누락된 경우**의 추정 결과

$$\widehat{FAMINC}_i = -7755 + 3211\,HEDU_i + 4777\,WEDU_i - 14311\,KL6_i$$

| | | | | |
|---|---|---|---|---|
| (se) | (11163) | (797) | (1061) | (5004) |
| ($p$-value) | (.488) | (.000) | (.000) | (.004) |

- 남편이 교육을 1년 더 받을 경우 연간 가계소득은 $3,211 증가

*HEDU*와 *WEDU* 추정치에 별 변화 없음 (불편추정량일 가능성 높음)

- *KL6*가 *HEDU* 및 *WEDU*와 독립이면, 누락되어도 전혀 영향 주지 않음

*KL6*은 6세 미만 자녀수
(남편 및 아내의 교육 수준과 무관)

## 6.6.2 종속변수와 관련되지 않은 변수가 포함되었을 때의 문제

- 올바른 모형이 다음과 같다고 하자

$$W_t = \beta_1 + \beta_2 E_t + \beta_3 M_t + e_t$$

- 부적절한 아래 모형을 추정하였다고 하자

$$W_t = \beta_1 + \beta_2 E_t + \beta_3 M_t + \beta_4 C_t + e_t$$

- 문제점

- $\beta_1, \beta_2, \beta_3$ 의 OLS 추정량은 불편추정량이기는 하지만, 분산이 증가하게 됨 ( $C$가 $E$ 및 $M$과 완전독립이 아니면)

⇒ 나머지 추정치의 정확도를 떨어뜨리고, 유의하지 않은 것으로 오해하게 만들 수 있음

- **가계소득 추정 결과**

$$\widehat{FAMINC}_i = -7755 + 3211 HEDU_i + 4777 WEDU_i - 14311 KL6_i$$

| | | | | |
|---|---|---|---|---|
| (se) | (11163) | (797) | (1061) | (5004) |
| ($p$-value) | (.488) | (.000) | (.000) | (.004) |

- 남편이 교육을 1년 더 받을 경우 연간 가계소득은 $3,132 증가

- **종속변수와 인위적인(무관한) 변수**($X5$, $X6$)가 **포함된 경우**의 추정 결과

$$\widehat{FAMINC}_i = -7759 + 3340\, HEDU_i + 5869 WEDU_i - 14200\, KL6_i + 889 X_{i5} - 1067 X_{i6}$$

| | | | | | | |
|---|---|---|---|---|---|---|
| (se) | (11195) | (1250) | (2278) | (5044) | (2242) | (1982) |
| ($p$-value) | (.500) | (.008) | (.010) | (.005) | (.692) | (.591) |

- $HEDU$와 $WEDU$ 추정치에 큰 변화 없음 (불편추정량일 가능성 있음)
- 주요 추정치의 표준오차가 크게 증가하는 문제 확인됨, 정확성 감소

$X5$, $X6$는 난수 발생시켜 만든 변수 (무관, 독립)

## 6.6.3 RESET 검정 (*Regression Specification Error Test*)

- 적절한 변수 누락 여부와 함수형태의 잘못된 설정을 검정

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + e_t \qquad \hat{Y}_t = b_1 + b_2 X_{2t} + b_3 X_{3t}$$

- 다음과 같은 두 개의 인위적인 모형을 생각해보자

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \gamma_1 \hat{Y}_t^2 + e_t$$

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \gamma_1 \hat{Y}_t^2 + \gamma_2 \hat{Y}_t^3 + e_t$$

- $\gamma_1$ 이나 $\gamma_2$ 가 0이 아니라면, **비선형함수**가 적절하다는 의미

- 따라서 RESET 검정의 가설은 아래와 같음

$$H_0 : \gamma_1 = 0, \quad H_1 : \gamma_1 \neq 0$$

$$H_0 : \gamma_1 = \gamma_2 = 0, \quad H_1 : \gamma_1 \neq 0 \ \ or \ \ \gamma_2 \neq 0$$

## ▪ RESET 검정 결과

- 귀무가설 기각되면,
  원래 모형이 부적절하여 향상될 수 있다는 의미
- 귀무가설을 기각하는데 실패할 경우,
  RESET 검정은 모형 설정의 오류를 알아낼 수 없다는 의미

```
proc autoreg ;
loglog:  model lq = lpb lpl lpr lm / reset ;
liner:      model q = pb pl pr m / reset ;
```

- **Ramsey (1969)의 RESET 검정 예: 맥주 수요모형**

<log-log model>

$$\ln(q_t) = \beta_1 + \beta_2 \ln(p_{Bt}) + \beta_3 \ln(p_{Lt}) + \beta_4 \ln(p_{Rt}) + \beta_5 \ln(m_t) + e_t$$

<linear model>

$$q_t = \beta_1 + \beta_2 p_{Bt} + \beta_3 p_{Lt} + \beta_4 p_{Rt} + \beta_5 m_t + e_t$$

| Ramsey RESET Test: LOGLOG Model | | | |
|---|---|---|---|
| F-statistic (1 term) | 0.0075 | Probability | 0.9319 |
| F-statistic (2 terms) | 0.3581 | Probability | 0.7028 |
| Ramsey RESET Test: LINEAR Model | | | |
| F-statistic (1 term) | 8.8377 | Probability | 0.0066 |
| F-statistic (2 terms) | 4.7618 | Probability | 0.0186 |

- **맥주 수요모형 RESET 검정 결과**

(1 term) : 예측치의 제곱변수가 추가된 경우

(2 terms) : 예측치의 제곱 및 세제곱 변수가 함께 추가된 경우

<log-log model>

- F-값이 아주 작음(0.0075, 0.3581), 이것들에 상응하는 p-값은 0.9319 및 0.7028로 유의수준(0.05)보다 훨씬 큼
  ⇒ log-log model이 부적절한 모형이라고 주장할 수 없음

<linear model>

- p-값은 0.0066 및 0.0186로 유의수준(0.05)보다 훨씬 작음
  ⇒ linear model은 부적절한 모형이라는 결론

| | |
|---|---|
| Model | loglog |
| Dependent Variable | lq |

Ordinary Least Squares Estimates

| | | | |
|---|---|---|---|
| SSE | 0.08991989 | DFE | 25 |
| MSE | 0.00360 | Root MSE | 0.05997 |
| SBC | -72.158707 | AIC | -79.164694 |
| Regress R-Square | 0.8254 | Total R-Square | 0.8254 |
| Durbin-Watson | 2.6306 | | |

Ramsey's RESET Test

| Power | RESET | Pr > F |
|---|---|---|
| 2 | 0.0074 | 0.9319 |
| 3 | 0.3581 | 0.7028 |
| 4 | 0.7385 | 0.5403 |

| Variable | DF | Estimate | Standard Error | t Value | Approx Pr > \|t\| |
|---|---|---|---|---|---|
| Intercept | 1 | -3.2432 | 3.7430 | -0.87 | 0.3945 |
| lpb | 1 | -1.0204 | 0.2390 | -4.27 | 0.0002 |
| lpl | 1 | -0.5829 | 0.5602 | -1.04 | 0.3080 |
| lpr | 1 | 0.2095 | 0.0797 | 2.63 | 0.0144 |
| lm | 1 | 0.9229 | 0.4155 | 2.22 | 0.0356 |

| | |
|---|---|
| Model | liner |
| Dependent Variable | q |

Ordinary Least Squares Estimates

| | | | |
|---|---|---|---|
| SSE | 318.483105 | DFE | 25 |
| MSE | 12.73932 | Root MSE | 3.56922 |
| SBC | 173.01346 | AIC | 166.007473 |
| Regress R-Square | 0.8221 | Total R-Square | 0.8221 |
| Durbin-Watson | 2.5088 | | |

Ramsey's RESET Test

| Power | RESET | Pr > F |
|---|---|---|
| 2 | 8.8377 | 0.0066 |
| 3 | 4.7618 | 0.0186 |
| 4 | 3.3019 | 0.0392 |

| Variable | DF | Estimate | Standard Error | t Value | Approx Pr > \|t\| |
|---|---|---|---|---|---|
| Intercept | 1 | 82.1587 | 17.9618 | 4.57 | 0.0001 |
| pb | 1 | -23.7426 | 5.4294 | -4.37 | 0.0002 |
| pl | 1 | -4.0774 | 3.8905 | -1.05 | 0.3046 |
| pr | 1 | 12.9243 | 4.1639 | 3.10 | 0.0047 |
| m | 1 | 0.001995 | 0.000776 | 2.57 | 0.0165 |

## 6.7 다중공선성 (multicollinearity)

- 설명변수들 사이에는 밀접한 관련이 있을 수 있음

  (두 변수가 공선(collinear)되어 있다고 라고 말함)

  ⇒ 이 경우 설명변수가 종속변수에 미치는 영향을 추정하기 어려움

  설명변수 자료는 풍부한 정보를 가지지 못함(poor data)

  (공선성 문제 collinearity problem 라고 말함)

(예1) 신문광고와 우대권(할인권) 나누어주기를 **동시에** 하는 경우

⇒ 매출액의 증가에 미치는 광고의 효과와 우대권의 효과를 **분리하여 추정**하기 어려움

(예2) 노동과 자본이 **고정된 비율**로 생산에 투입되는 경우(버스-운전사)

⇒ 노동과 자본의 한계생산물을 각각 분리하여 추정하기 어려움

## 6.7.1 다중공선성이 유발하는 효과들

- $R^2$ 및 모형의 $F$ 값이 높음에도 불구하고 추정치들의 t-값이 낮게 나타남, 추정치들이 유의하지 않은 것으로 오해하게 함
- 몇 개의 표본을 추가하거나 제외시키면,
  추정치들이 매우 민감하게 달라짐
- 모형에서 명백하게 유의하지 않은 설명변수를 삭제하면
  추정치들이 민감하게 달라짐

❖ 공선 관계가 표본외 관찰값에도 동일하게 존재하는 경우에는 예측의 정확도가 높을 수 있음

- **다중공선성 문제 (**multicollinearity problem)

- 설명변수들 사이에 **정확한** 선형관계가 존재하는 경우,

  OLS를 이용해서는 모수 추정치를 구할 수 없음

$$\mathrm{var}(b_2) = \frac{\sigma^2}{(1-r_{23}^2)\sum_{i=1}^{N}(x_{i2}-\bar{x}_2)^2}$$

- 설명변수들 사이에 **거의 정확한** 선형관계가 존재하는 경우,

  OLS 추정량의 일부 분산, 표준오차가 큰 값으로 나타남

⇒ 구간추정이 넓은 폭으로 나타나 모수의 참값에 대한 정보를 얻기 어려움

## 6.7.2 자동차 연비 사례

*MPG* = 연비 (miles per gallon)

*CYL* = 실린더의 수 (number of cylinders; V4, V6, V8,…)

*ENG* = 배기량 (engine displacement in cubic inches)

*WGT* = 차량 무게 (vehicle weight in pounds)

- ***CYL, ENG, WGT*가 *MPG*에 미치는 영향을 추정하려고 함**
- 일반적으로 대형차(소형차)의 경우 *CYL*, *ENG*, *WGT*가 모두 큼(작음)

  ⇒ 설명변수들 사이에 다중공선성이 존재할 가능성이 높음

## ▪ 자동차 연비 사례 추정결과

- 설명변수가 하나만 포함된 모형의 추정결과

$$\widehat{MPG}_i = 42.9 - 3.558\,CYL_i$$
$$(\text{se}) \quad (0.83) \quad (0.146)$$
$$(p\text{-value}) \quad (0.000) \quad (0.000)$$

- 설명변수가 모두 포함된 모형의 추정결과

$$\widehat{MPG}_i = 44.4 - 0.268\,CYL_i - 0.0127\,ENG_i - 0.00571\,WGT_i$$
$$(\text{se}) \quad (1.5) \quad (0.413) \quad (0.0083) \quad (0.0071)$$
$$(p\text{-value}) \quad (0.000) \quad (0.517) \quad (0.125) \quad (0.000)$$

✓ $CYL$ 변수 추정치의 유의성이 크게 낮게 나타나고 있음

## 6.7.3 다중공선성의 탐지 및 해결

- **다중공선성 탐지 방법**

(1) 설명변수들 사이의 상관관계 측정 (상관계수가 0.8 이상)

(2) $R^2$ 나 F-값이 높음에도 불구하고, 추정치의 t-값이 낮을 때

(3) 다음과 같은 보조적인 회귀모형 추정, $R^2$ 가 0.8 이상

$$x_{i2} = a_1 x_{i1} + a_3 x_{i3} + \cdots + a_K x_{iK} + error$$

(4) 표본을 몇 개 늘리거나 줄이면, 추정치 크기가 크게 변화

## ▪ 다중공선성 해결 방법

- 상관관계 높은 설명변수 삭제
  - 설명변수들 사이의 상관계수 계산해서 찾음
- 표본의 수 늘림
- 비표본정보의 이용, 제약 부과

## 6.8 예측

$$y_i = \beta_1 + x_{i2}\beta_2 + x_{i3}\beta_3 + e_i$$

- 단순회귀 경우와 유사함
- $(1,\ x_{02},\ x_{03})$ 에서의 점예측치는 아래와 같음

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

**<과제>**

**6.6**

**6.10** (beer_new.dat 자료를 이용하시오)

**Eviews output을 출력하고,**

**출력물의 빈 여백에 간단하게 답을 적으시오.**

※ 참고: 6.10 문제에 필요한 data는 사이버강의실에 있음

http://principlesofeconometrics.com/