제9장 다중회귀 추정량의 성질

9.1 모형의 구성항목들에 대한 가정

- X에 관한 가정
  - 설명변수 표본값 고정
  - 비특이성: 설명변수들의 관측값들 간에 선형종속의 관계가 존재하지 않음
- u에 관한 가정
  - 오차평균0
  - 동일분산
  - 독립추출
  - 정규분포

제9장 다중회귀 추정량의 성질

9.2 OLS 추정량의 평균

- X에 관한 두 가정 + ‘오차평균 0’이 성립하면 OLS 추정량은 비편향
- 증명: $\hat{\beta}_j = \beta_j + \frac{\sum_{i=1}^{n} \hat{r}_{ij} u_i}{\sum_{i=1}^{n} \hat{r}_{ij}^2}$을 보인 다음 양변에 평균을 취하면 오차평균 0 가정으로 인해 둘째 항은 0이 되고 $E(\hat{\beta}_j) = \beta_j$가 됨

제9장 다중회귀 추정량의 성질

9.3 변수를 누락시키면 어떻게 될까

- 관심을 갖는 모형은 $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$. 즉, X_2를 통제하고자 함.
- 그런데 (무슨 이유에서든) Y를 X_1에 대하여만 회귀함(단순회귀) => $\tilde{\beta}_0$, $\tilde{\beta}_1$
- Y를 X_1과 X_2에 대하여 (제대로) 회귀하여 구한 OLS 추정량을 $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$이라 하자.
- 그러면 다음 항등식이 성립함

$$\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2 \tilde{\delta}_1$$

  - 여기서 $\tilde{\delta}_1$은 X_2를 X_1에 대하여 (절편을 포함하여) 회귀할 때의 기울기 계수
- 우변은 'X_2를 통제할 때 X_1 변화의 효과'와 'X_1이 변할 때 X_2가 변하는 정도($\tilde{\delta}_1$)에 X_1을 통제할 때 X_2 변화의 효과'를 합한 것

총효과 = 직접효과 + 간접효과

제9장 다중회귀 추정량의 성질

9.3 변수를 누락시키면 어떻게 될까 (계속)

- 그러므로 $E(\tilde{\beta}_1) = \beta_1 + \beta_2\tilde{\delta}_1$
- X_2를 통제하고자 하지만 통제하지 않아도 편향되지 않는 경우는: (i) X_1을 통제하면 X_2가 Y에 영향을 미치지 않는 경우($\beta_2 = 0$); (ii) X_1과 X_2가 상관되지 않은 경우.
  - β_2가 0이 아니더라도 X_1과 X_2 간의 상관이 작으면 편향은 작을 것임
- Y에 영향을 미치면서 X_1과 상관된 통제변수는 누락시키지 말아야 함
- 반대로, X_2를 통제하고 싶지 않은데 우변에 X_2를 포함시켜서 회귀할 경우에도, X_2가 Y에 별도의 영향을 미치지 않거나(즉, u와 비상관) X_2가 X_1과 비상관이면 편향(bias)을 야기하지 않음
  - 하지만 X_1과도 상관되고 u와도 상관되는 변수를 우변에 추가로 통제하면 편향이 초래됨

제9장 다중회귀 추정량의 성질

9.4 OLS 추정량의 분산

- X에 관한 두 가정 + 오차평균 0 + 동일분산 + 독립추출 가정하에서 다중회귀 모형의 OLS 추정량의 분산을 구하면 다음과 같음

$$var(\hat{\beta}_j) = \frac{\sigma^2}{\sum_{i=1}^{n} \hat{r}_{ij}^2}$$

  - 증명은 단순회귀의 경우와 유사

- 그런데 분모는 X_j를 여타 설명변수들에 대하여 회귀할 때의 SSR이므로 SSR_j라 표기하고, 그 회귀에서 총제곱합을 SST_j, R제곱을 R_j^2이라 하면, $SSR_j = SST_j(1 - R_j^2)$이 성립하므로

$$var(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1 - R_j^2)} = \frac{\sigma^2}{SST_j} \cdot \frac{1}{1 - R_j^2}$$

단순회귀 시 분산

"분산팽창계수"

제9장 다중회귀 추정량의 성질

9.5 가우스 마코프 정리

- 가우스 마코프 정리(Gauss Markov Theorem): X에 관한 두 가정 + 오차평균 0 + 동일분산 + 독립추출 가정("가우스 마코프 가정"이라 함)하에서 OLS는 BLUE (가장 좋은 선형 비편향 추정량)
  - 동일분산과 독립추출이 중요함
  - 이분산(heteroscedasticity)이나 자기상관(serial correaltion)이 있으면 OLS는 BLUE가 아님
- 단순모형과 달리 다중회귀에서는 계수가 여럿이며, 가우스 마코프 정리는 "계수들의 어떤 선형결합을 고려하더라도 OLS의 분산이 선형 비편향 추정량 중에서는 가장 작다"는 것을 의미함
  - 예를 들어 $\hat{\beta}_j$가 OLS 추정량이고 $\tilde{\beta}_j$가 어떤 선형 비편향 추정량이라 하면 $var(\hat{\beta}_1 - 2\hat{\beta}_2) \leq var(\tilde{\beta}_1 - 2\tilde{\beta}_2)$
  - 다른 어떠한 계수들의 결합(선형 결합)을 고려하더라도 부등호는 마찬가지

제9장 다중회귀 추정량의 성질

9.6 설명변수의 추가 또는 누락과 추정량의 분산

- 앞에서 모형이 $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$일 때, $\beta_2 = 0$이거나 X_1과 X_2가 비상관이면 X_2를 누락시켜도 편향이 야기되지 않는다고 하였음
- 만약 $\beta_2 = 0$이고 X_1과 X_2가 상관되면, 단순회귀 모형과 다중회귀 모형에서 오차항은 모두 u이므로
  - 단순회귀 β_1 계수추정량의 분산 = $\frac{\sigma^2}{SST_1}$, 다중회귀 β_1 계수추정량의 분산 = $\frac{\sigma^2}{SST_1} \times \frac{1}{1-R_1^2}$
  - $R_1^2 > 0$일 것이므로, 단순회귀 계수추정량이 분산이 더 작음
- 만약 $\beta_2 \neq 0$이고 X_1과 X_2가 비상관이면, 단순회귀에서 오차항은 $v = u + \beta_2 [X_2 -$

제9장 다중회귀 추정량의 성질

9.6 설명변수의 추가 또는 누락과 추정량의 분산(계속)

- 요약하면 다음과 같음
- Y에 별도의 영향을 미치지 않으면서 X_1과 연관된 변수를 우변에 추가하면, X_1 변수 내의 정보가 삭감되어 X_1 계수 추정량의 표집분산이 커지고 정확도가 떨어짐
- Y에 대한 별도의 설명력을 가지면서도 X_1과 상관되지 않은 변수를 통제하면 설명불가 요인들(오차항)의 변동성이 줄어들면서도 X_1 내의 정보가 삭감되지 않아 다중회귀가 단순회귀보다 더 효율적인 추정량을 제공함

제9장 다중회귀 추정량의 성질

9.7 OLS 추정량의 분산의 추정과 표준오차

- 단순회귀의 경우와 유사함
- 오차항의 분산인 σ^2은 $s^2 = \frac{1}{df}\sum_{i=1}^{n}\hat{u}_i^2 = \frac{SSR}{df}$에 의하여 추정함. 이때 $df = n - k - 1$
  - s^2을 '회귀의 표준오차'라 함
- OLS 계수 추정량의 분산식에서 σ^2을 s^2으로 치환하여 분산을 추정할 수 있음 $\widehat{var}(\hat{\beta}_j) = \frac{s^2}{SSR_j}$.
- 여기에 제곱근을 취한 것이 $\hat{\beta}_j$의 통상적인 '표준오차'(standard error)

$$se(\hat{\beta}_j) = \frac{s}{\sqrt{SSR_j}}$$

제9장 다중회귀 추정량의 성질

9.8 OLS 추정량의 표집분포

- 단순회귀의 경우와 유사함
- 설명변수 표본값 고정, 비특이성, 오차평균0, 동일분산, 독립추출, 정규분포 가정하에서 $\hat{\beta}_j$는 정규분포를 가짐
  - 평균과 분산은 이미 구하였음
- 그뿐 아니라, $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$의 어떤 선형결합이든지 정규분포를 가짐
  - 즉, 어떤 $\lambda_0, \lambda_1, \ \ldots, \lambda_k$에 대해서든(nonrandom할 것), $\lambda_0\hat{\beta}_0 + \lambda_1\hat{\beta}_1 + \cdots + \lambda_k\hat{\beta}_k$은 정규분포를 가짐
  - 평균과 분산은 어렵지 않게 구할 수 있음
  - 예를 들어 $\hat{\beta}_1 - \hat{\beta}_2$는 정규분포를 가짐

제9장 다중회귀 추정량의 성질

9.9 신뢰구간

- 단순회귀의 경우와 유사함
- 신뢰구간의 양끝값은 추정값 ± 임계값 × 표준오차
  - 표준오차는 t_{df} 분포로부터 구함$(df = n - k - 1)$
- 계수들의 선형결합의 신뢰구간도 표준오차만 구하면 이와 동일한 방식을 이용하여 구할 수 있음
- 하지만 이 경우 곧이곧대로 표준오차를 계산하기 복잡하며, '모수변환' 방법을 이용하여 간략히 계산을 수행할 수 있음

제9장 다중회귀 추정량의 성질

9.9 신뢰구간 (계속)

- 예: $\log(price) = \beta_0 + \beta_1 \log(lotsize) + \beta_2 bedrooms + u$에서 $lotsize = 5000$이고 $bedrooms = 3$인 집들의 평균 로그 가격(θ)의 신뢰구간
- $\theta = \beta_0 + \beta_1 \log(5000) + \beta_2 \times 3$
- $\beta_0 = \theta - \beta_1 \log(5000) + \beta_2 \times 3$을 대입하면 모형은 다음이 됨
  $$\log(price) = \theta + \beta_1[\log(lotsize) - \log(5000)] + \beta_2(bedrooms - 3) + u$$
- 그러므로 $\log(price)$를 $\log(lotsize) - \log(5000)$과 $bedrooms - 3$에 대하여 회귀할 때의 절편이 바로 θ
- R을 사용한다면 lm(log(price)~I(log(lotsize)-log(5000))+I(bedrooms-3))을 이용

```
> library(Ecdat)
> data(Housing)
> ols <- lm(log(price)~I(log(lotsize)-log(5000))+I(bedrooms-3),
  data=Housing)
> summary(ols)

Call:
lm(formula = log(price) ~ I(log(lotsize) - log(5000)) + I(bedrooms -
    3), data = Housing)

Residuals:
     Min       1Q   Median       3Q      Max
-0.95071 -0.17457  0.

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                 11.08939    0.01227 903.860   <2e-16 ***
I(log(lotsize) - log(5000))  0.50151    0.03095  16.201   <2e-16 ***
I(bedrooms - 3)              0.14587    0.01670   8.733   <2e-16 ***
```

여기 추정값과 표준오차

질

$+ u$에서 $lotsize = 5000$이고

간

다음이 됨

$\beta_2(bedrooms - 3) + u$

$edrooms - 3$에 대하여 회귀할

00))+I(bedrooms-3))을 이용

```
> library(Ecdat)
> data(Housing)
> ols <- lm(log(price)~I(log(lotsize)-log(5000))+I(bedrooms-3),
  data=Housing)
> summary(ols)

Call:
lm(formula = log(price) ~ I(log(lotsize) - log(5000)) + I(bedrooms -
    3), data = Housing)

Residuals:
     Min       1Q   Median       3Q      Max
-0.95071 -0.17457  0.

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                 11.08939    0.01227 903.860   <2e-16 ***
I(log(lotsize) - log(5000))  0.50151    0.03095  16.201   <2e-16 ***
I(bedrooms -
```

여기 추정값과 표준오차

```
> confint(ols)
                                2.5 %     97.5 %
(Intercept)                 11.0652938 11.1134946
I(log(lotsize) - log(5000))  0.4407007  0.5623103
I(bedrooms - 3)              0.1130589  0.1786842
```

95% 신뢰구간

질

$+ u$에서 $lotsize = 5000$이고

간

다음이 됨

$\beta_2(bedrooms - 3) + u$

$edrooms - 3$에 대하여 회귀할

이용