

2022년도 하반기 외국어 및 종합시험 답안지

학과별	과정별	시험 과 목 명	수험번호	성 명	감독위원확인
통계	박사	선형모형특강	5250	황성원	

1. A: $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_g x_{gi} + \epsilon_{1i}$

B: $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_g x_{gi} + \dots + \beta_p x_{pi} + \epsilon_{2i}, i=1, \dots, n, (p > g)$

(1-1) model A $\Rightarrow H_0: \beta_j = 0$ vs $H_1: \beta_j \neq 0$

$$T = \frac{\hat{\beta}_j - 0}{s.e(\hat{\beta}_j)} \sim t(n-g-1) \text{ 임을 이용하여 } (\epsilon_{1i} \overset{H_0}{\sim} N(0, \sigma^2) \text{ 가짐})$$

유의수준 α 에서의 p-value 를 계산하고 H_0 의 기각여부를 판단한다.

(1-2) model A $\Rightarrow H_0: \beta_1 = \beta_2 = \dots = \beta_g = 0$ vs $H_1: \text{not } H_0$

ANOVA Table

	SS	df	MS	F_0
Regression	SSR	g	MSR	$\frac{MSR}{MSE}$
Error	SSE	$n-g-1$	MSE	
Total	SST	$n-1$		

$$\Rightarrow F_0 = \frac{MSR}{MSE} \overset{H_0}{\sim} F_{g, n-g-1} \text{ 을}$$

이용하여 유의수준 α 에서의 p-value 를 계산하고

H_0 의 기각여부를 판단한다.

(1-3) $H_0: \beta_{g+1} = \beta_{g+2} = \dots = \beta_p = 0$ vs $H_1: \text{not } H_0$

Partial F-test $\Rightarrow \text{Reduced: } y_i = \beta_0 + \beta_1 x_{1i} + \dots + \beta_g x_{gi} / \text{Full: } y_i = \beta_0 + \beta_1 x_{1i} + \dots + \beta_p x_{pi}$

$$F_0 = \frac{(SSE_{\text{Reduced}} - SSE_{\text{Full}}) / (p-g)}{SSE_{\text{Full}} / (n-p-1)} \overset{H_0}{\sim} F_{p-g, n-p-1} \text{ 임을 이용하여}$$

유의수준 α 에서의 p-value 계산후

H_0 의 기각여부 판단,

<서식6>

2022년도 하반기 외국어 및 종합시험 답안지

학과별	과정별	시험 과 목 명	수험번호	성 명	감독위원확인
통계	박사	선형모형특강	5250	황성운	

(1-4) model A를 Reduced로, model B를 Full로 둘 수 있으므로

(1-3)에서 기술한 절차를 그대로 따라가면서 검토하면 된다,

(1-5) (1-3)과 (1-4) 모두 Reduced model과 Full model이
각각 같은 경우이므로 결국 같은 절차이다.

2. (2-1) 중회귀모형 $y = X\beta + \epsilon$ 에서 변수 x_j 가 $x_j' = bx_j$ ($b > 0$)으로 대체된다면

모형이 $y = \beta_0 + \beta_1 x_1 + \dots + \beta_j' \cdot bx_j + \dots + \beta_k x_k + \epsilon$ 로 바뀌게 된다.

이에 대한 β_j' 의 최소제곱추정량은 $\hat{\beta}_j' = \frac{1}{b} \hat{\beta}_j$ 이 된다.

(2-2) (2-1)에서 언급된 스케일링은 수치의 크기에만 영향을 줄 뿐
모형 전체의 해석에는 큰 영향을 주지 않는다, 그 이유는 최소제곱추정량이
 $\frac{1}{b}$ 의 상수배 정도의 크기 변화만 일어났기 때문이다.

3. 다중공선성 : 중회귀모형에서 설명변수들 사이에 연관성이 존재하여
추정량과 분산이 과도하게 커지는 현상이다,

해결책 : 서로 연관성이 있다고 여겨지는 변수들 중 하나를 제거,
RIDGE regression, LASSO regression 등의 방법으로
penalty를 주어서 추정량을 계산.

<서식7>

2022년도 하반기 외국어 및 종합시험 답안지

학과별	과정별	시험 과 목 명	수험번호	성 명	감독위원확인
통계	박사	선형모형특강	5250	황성윤	

4. 중회귀모형이 $\hat{y} = X\beta + \varepsilon$ 과 같을때

LASSO regression에서는 다음과 같은 방법을 통해 $\hat{\beta}$ 을 구한다.

$$\arg \min_{\beta} (\hat{y} - X\beta)^T (\hat{y} - X\beta) + \lambda \|\beta\|, \quad (\lambda: \text{penalty를 조절해주는 상수})$$

즉 LASSO regression에서는 강한 penalty를 통해

회귀계수의 크기를 큰 쪽으로 축소시켜서 일부의 회귀계수 추정량을

0으로 수렴시킨다는 특징이 있다, 이 때문에 LASSO regression에서는

추정량이 Sparse 해지게 되고 이를 통해 설명변수의 개수가 많은 경우

변수선택을 할 수 있다는 특징이 있다.