데이K0004-01 회귀분석 중간고사

2020년 10월 26일(월)

모든 문제에 대해 R을 이용하여 다음의 문제들을 풀기 바랍니다. (102점 만점)

1. 다음은 몸무게(X kg)를 가지고 심장 박동율(Y)을 회귀모형을 통해 예측하기 위해 얻어진 6명의 사람들로부터 얻어진 요약된 자료이다. (33점)

$$\sum_{i=1}^{6} X_i = 488, \quad \sum_{i=1}^{6} Y_i = 319, \quad \sum_{i=1}^{6} X_i^2 = 40092, \quad \sum_{i=1}^{6} Y_i^2 = 17399, \quad \sum_{i=1}^{6} X_i Y_i = 26184$$

위의 요약된 자료를 이용하여 다음의 문항에 답하시오. (계산은 반드시 R프로그램을 이용하여 구하시오).

(1) X와 Y의 공분산을 구하고 그 의미를 설명하시오. (3점)

(2) X와 Y의 상관계수를 구하고 그 의미를 설명하시오. (3점)

(3) 위 자료에 대한 회귀식에 대한 $\hat{\beta}_0$ and $\hat{\beta}_1$을 구하고 추정된 회귀식을 제시하시오. (3점)

(4) 위에서 구한 모형의 결정계수(R^2)를 구하시오.

(5) MSE(Mean Square Error)를 구하고 그 의미를 설명하시오. (3점)

(6) 최소제곱에 의해 구한 σ^2의 추정치 $\hat{\sigma}^2$와 정규분포 하에서의 최대가능도 추정에 의해 구한 $\tilde{\sigma}^2$를 구하시오. 이 둘의 차이에 대해 설명하시오.

(7) β_1에 대한 95% 신뢰구간을 구하고 해석하시오. (3점)

(8) 유의수준 $\alpha = 0.05$ 에서 $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$에 대한 t-검정을 하시오. 이 결과는 (7)에서 구한 신뢰구간의 결과와 일치되는지 설명하시오? (3점)

(9) $X = 88$일 때 Y의 평균에 대한 점 추정치 $\widehat{E(Y)}$를 구하시오. 그리고 이 추정치에 대한 신뢰구간을 구하고 해석하시오. (3점)

(10) 어떤 특정한 개인이 몸무게가 88 kg인 사람의 심장박동율을 예측하고 95% 신뢰구간을 구하시오. 이것은 (9)에서 구한 추정치와 무엇이 다른지를 비교 설명하시오. (3점)

(11) $\hat{Y}$ 의 가장 작은 분산을 주는 측정된 X 값은 무엇인가? 근거에 기반하여 이유를 설명하시오. (3점)

2. 다음 표는 수용액에서의 농도(X)와 색채측정기로 측정한 채도값(Y)이다. (30점)

X	Y
40	69
50	175
60	272
70	335
80	490
90	415
40	72
60	265
80	492
50	180

(1) (X, Y) 산점도를 그리고 상관계수를 구하여 수용액에서의 농도와 채도값과의 관계에 대해서 설명하시오. (3점)

(2) 통계적 모형 $Y=\beta_0+\beta_1 X+\epsilon$에 적합하고 회귀식을 표현하시오. (1)의 산점도에 회귀선을 추가하시오 (3점)

(3) $H_0 : \beta_1 = 0$, $H_1 : \beta_1 \neq 0$에 대한 유의수준 5%에서 검정하시오. (3점)

(4) 결정계수(R^2)를 구하고 해석하시오. (1)에서의 구한 상관계수값과 결정계수를 비교하여 상관계수와 결정계수의 관계를 설명하시오. (3점)

(5) (2)번의 모형을 이용하여 각 X에 대응하는 $\hat{Y}$을 추정하시오. (3점)

(6) ($\hat{Y}, \hat{\epsilon}$) 그림을 그리고 독립성과 등분산성에 대해 설명하시오. (3점)

(7) 잔차가 정규분포를 따른다고 할 수 있는지 Q-Q plot과 정규성검정을 사용하여 설명하시오. (3점)

(8) X의 평균값에서 Y의 추정값과 신뢰구간을 구하시오. (3점)

(9) X=(45, 55, 65)에서 Y의 추정값과 신뢰구간을 구하시오. (3점)

(10) R 행렬 프로그램을 이용하여 (2)의 회귀식의 추정계수값과 MSE, 결정계수(R^2)를 구하시오. (3점)

3. 문제에 주어진 SAT성적 데이터(엑셀제공)는 A 대학 재학생과 관련한 것이다. 각 학생에 대한 측정된 4개 변수는 다음과 같다. (39점)

Y : College GPA

X_1 : High school GPA

X_2 : SAT Total

X_3 : Quality of letters of recommendation

(1) (X_1, Y), (X_2, Y), (X_3, Y)에 대해 각각 산점도를 그리시오. 상관계수를 구하여 각 설명변수들과 반응변수인 대학 GPA와의 관계를 설명하시오. (3점)

(2) X_1, X_2, X_3간의 산점도를 그리고 설명변수들간의 상관계수를 통해 설명변수들간의 관례를 설명하고 다중회귀모형에 미칠 수 있는 영향을 제시하시오. (3점)

(3) 통계적 모형 $Y=\beta_0+\beta_1 X_1+\beta_2 X_2+\beta_3 X_3+\epsilon$ 을 적합하여 제시하시오. (3점)

(4) $H_0 : \beta_1 = 0$, $H_1 : \beta_1 \neq 0$에 대한 t-통계량과 p-값을 구하고 유의수준 0.05에서 검정하시오. (3점)

(5) $H_0 : \beta_2 = 0$, $H_1 : \beta_2 \neq 0$에 대한 t-통계량과 p-값을 구하고 유의수준 0.05에서 검정하시오. (3점)

(6) $H_0 : \beta_3 = 0$, $H_1 : \beta_3 \neq 0$에 대한 t-통계량과 p-값을 구하고 유의수준 0.05에서 검정하시오. (3점)

(7) ($\hat{Y}, \hat{\epsilon}$) 그림을 그리고 잔차의 독립성과 등분산성에 대해 설명하시오. (3점)

(8) 잔차의 독립성검정을 하시오. (3점)

(9) 잔차가 정규분포를 따른다고 할 수 있는지 Q-Q plot과 정규성검정을 사용하여 설명하시오. (3점)

(10) 반응변수인 대학 GPA에 가장 영향을 주는 설명변수는 무엇인가? 객관적인 근거를 기반으로 설명하시오. (3점)

(11) 회귀모형에 대한 결정계수(R^2)를 구하고 그 의미를 설명하시오. (3점)

(12) $X_1 = 3.0$, $X_2 = 1200$, $X_3 = 6$ 일 때 예측되는 대학 GPA의 추정값과 신뢰구간은 어떻게 되는가? (3점)

(13) R 행렬 프로그램을 이용하여 (3)의 회귀식의 추정계수값과 MSE, 결정계수(R^2)를 구하시오. (3점)