회귀분석 시험 1

2023년 4월 18일

*학번:
*이름:
*주의
- 계산결과를 구할 필요는 없으며 계산식만 제시할 것 (예. (38.5-20.3)/4 등)

It is not necessary to obtain the calculation results, only the calculation formula is presented (e.g. (38.5-20.3)/4, etc.)
- 답 작성할 공간이 부족할 시 뒷면을 사용할 것

Use the back when there is not enough space to fill out the answer

A. 다음 각 질문에 답하시오.

1. 어떤 자료에 대하여 정규성 검정을 시행한 경우에 검정통계량의 P-값이 작으면 자료가 정규분포를 따른다고 결론을 내릴 수 있는가? (2점)
2. 자유도(degrees of freedom)의 정의는? (2점)
3. Standardized residual $e_{sd,i}$ 와 Studentized residual e_{si}의 차이점을 기술하시오. (5점)
4. 결정계수(coefficient of determination, R^2) 대신에 수정결정계수(adjusted R^2)를 쓰는 이유는? (3점)
5. 회귀분석 모형에서 모수의 maximum likelihood estimator를 구하기위해서는 정규성 가정이 필요한가? (3점)
6. 두모집단의 평균을 비교하기위해 t-검정에서 두 그룹의 분산이 다른 경우도 회귀분석 모형으로 표현할 수 있는가? (3점)
7. 두 표본 자료의 정규성검정을 위해서는 전체 자료에 대해서 q-q plot을 그려서 확인하면 되는가? (3점)

B. Z_1, Z_2, Z_3, Z_4가 서로 독립이고 $N(0,1)$의 분포를 따른다고 가정하자. 다음 각 변수들의 분포를 명시하라. (분포의 자유도를 명확하기 표기할 것) (각 2점)

1. Z_1^2 2. $\frac{Z_1^2+Z_2^2}{Z_3^2+Z_4^2}$ 3. $\frac{Z_1}{|Z_2|}$ 4. $\frac{Z_1+Z_2}{\sqrt{Z_3^2+Z_4^2}}$ 5. Z_1^2/Z_2^2

C. 다음 모형 중에서 선형모형을 모두 고르시오. (3점)

1. $Y = \beta_0 + e^{\beta_1 X_1} + \epsilon$
2. $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \epsilon$
3. $Y = \beta_0 + \beta_1 \log(X_1) + \epsilon$
4. $Y = \beta_0 + e_1^{\beta} e_1^{X} + \epsilon$
5. $Y = \beta_0 + \beta_1 \log(X_1) + \beta_2 [\log(X_1)]^2 + \epsilon$

D. 다음 각 질문에 답하고 그 이유를 자세히 설명하시오. (각 5점)

1. X_1, X_2의 두 설명변수에 대해서 다중회귀분석을 적합한 경우에 X_1의 부호는 X_1하나만 포함된 단순회귀분석모형을 적합한 경우의 X_1 계수의 부호와 일치하는 경우와 일치하지 않은 경우의 예를 들어 보시오. (scatter plot 등)
2. Interaction effect가 유의한 회귀분석모형에서 X_1, X_2의 main effect (주효과)가 유의하지 않게 나오면 X_1, X_2은 Y에 대해 둘 다 유의한 marginal effect가 없다고 할 수 있는가?

E. 회귀분석을 수강한 학생들 남학생과 여학생 성적 차이가 있는지를 비교해 보고자 한다. 다음과 같이 성적에 대해 가정해 보자..

$$\begin{array}{lllll} \text{남학생} : & y_1, & \cdots, & y_{n_1} & \sim i.i.d.\ N(\mu_1,\sigma^2) \\ \text{여학생} : & y_{n_1+1}, & \cdots, & y_{n_1+n_2} & \sim i.i.d.\ N(\mu_2,\sigma^2) \end{array}$$

1. 이 자료에 대하여 다음과 같은 회귀분석 모형을 고려하였다. X_1은 성별을 나타내는 변수로서

$$X_1 = \begin{cases} 0, \text{여학생인 경우} \\ 1, \text{남학생인 경우} \end{cases}$$

이 자료에 대하여 모형

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,\ \epsilon_i \sim i.i.d. N(0,\sigma^2)$$
$$i = 1,\cdots,n_1 \text{ for 남학생}, \quad i = n_1+1,\cdots,n_1+n_2 \text{ for 여학생}$$

을 적합시켰다. 이 모형의 모수 β_0, β_1을 μ_1, μ_2 로 표시하시오. 또 모수 β_0, β_1의 최소제곱추정량을 구하시오. (8점)

2. $H_0 : \mu_1 = \mu_2$ 의 가설을 검정하기 위한 통계량과 $H_0 : \beta_1 = 0$ 가설을 검정하기 위한 통계량이 일치하는가? 구체적으로 설명하시오. (5점)

F. We want to derive the estimator of β in the model $Y = X\beta + \epsilon$ subject to $C\beta = 0$. (15 pt)

1. Show that the estimator of β subject to $C\beta = 0$ is $\hat{\beta}_C = \hat{\beta} - (X'X)^{-1}C'[C(X'X)^{-1}C']^{-1}C\hat{\beta}$, where $\hat{\beta} = (X'X)^{-1}X'Y$. Hint. Use a Lagrange multiplier λ and minimize $S = (Y - X\beta)'(Y - X\beta) + \lambda'(C\beta - 0)$ with respect to β and λ.

2. Show that $\hat{\beta}X'Y - \hat{\beta}_C X'Y = (C\hat{\beta})'[C(X'X)^{-1}C']^{-1}C\hat{\beta}$.

G. If $\boldsymbol{Y}$ is a random vector with $\boldsymbol{\mu}$ and variance $\boldsymbol{V}$, if $\boldsymbol{A}$ is a symmetric matrix of constants, then show the following. (7 pts)

$$E(\boldsymbol{Y}'\boldsymbol{A}\boldsymbol{Y}) = tr(\boldsymbol{A}\boldsymbol{V}) + \boldsymbol{\mu}'\boldsymbol{A}\boldsymbol{\mu}$$

H. 회귀모형 $\boldsymbol{Y}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\epsilon}$ 에 대하여 SSE를 이차형식으로 유도하고 그 기댓값을 구하시오. 힌트. G의 결과를 이용하시오. (8점)

I. A company offers a training course for the GRE. They give their students a test at the end of the course, graded from 0 to 100. They would like to use that test in the future to predict how well students will score on the GRE. They have scores on their test and the GRE for a sample of students. Thus, X = score on the company's test and Y = score on the GRE, which ranges from 100 to 320. They plan to use the usual simple linear regression model.

1. One of the company analysts states that the intercept should be fixed at 100, because that's the lowest the GRE can be. Suppose the intercept is set to 100 for this situation. Write the regression model.

2. Write the full and reduced models to test whether or not it makes sense to set the intercept to be 100.

3. Write the sum that is to be minimized to get the least squares regression line, if the model you wrote in Part (b) is used.

J. Suppose we have the following two multiple linear regression models

Model (1): $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$

Model (2): $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i$

where errors are $i.i.d. \sim N(0, \sigma^2)$ both models. We first perform the analysis for Model (1) in R:

```
> fit12 = lm(Y ~ X1 + X2)
> summary(fit12)
Call:
lm(formula = Y ~ X1 + X2)

Residuals:
    Min              1Q          Median      3Q          Max
    -1.72610    -0.71385        0.03204     0.62244     3.04545
Coefficients:
                Estimate        Std. Error      t value     Pr(>|t|)
(Intercept)     -0.002956       0.094429        -0.031      0.975
    X1          2.171693        0.108222        20.067      <2e-16
    X2          2.949736        0.098936        29.814      <2e-16
Residual standard error: 0.9428 on 97 degrees of freedom
Multiple R-squared: 0.938,              Adjusted R-squared: ???
F-statistic: 733.5 on 2 and 97 DF,      p-value: < 2.2e-16
```

1. Calculate the adjusted R-square value from the output. (Just write the equation using numbers in the output)

2. Calculate the SSR (Regression Sum of Squares) from the output.

Now we perform the analysis for Model (2) in R:

```
> fit = lm(Y ~ X1 + X2 + X3)
> summary(fit)
Call:
lm(formula = Y ~ X1 + X2 + X3)

Residuals:
    Min              1Q          Median      3Q          Max
-1.72110        -0.71459        0.02617   0.62992    3.04839
Coefficients:
                Estimate        Std. Error      t value     Pr(>|t|)
(Intercept)    -0.002549        0.095098        -0.027       0.979
    X1          2.184818        0.217837        10.030       <2e-16
    X2          2.968544        0.288143        10.302       <2e-16
    X3          -0.063097       0.907274        -0.070       0.945
Residual standard error: 0.9476 on 96 degrees of freedom
Multiple R-squared: 0.938,              Adjusted R-squared: 0.936
F-statistic: 484 on 3 and 96 DF,        p-value: < 2.2e-16
```

3. From the outputs of Models (1) and (2), perform the hypothesis test, $H_0 : \beta_1 = \beta_2 = 0$ *vs* $H_1 :$ not H_0. Write down the test method and calculate the test statistics.

4. From the outputs of Models (1) and (2), perform the test to determine whether addition of X3 in Model (1) is a statistically significant or not.