학번: 이름:

회귀분석 - 기말고사

1. 행렬과 벡터를 사용한 중선형 회귀모형은 다음과 같다.

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{e}$$

(a) 잔차벡터가 아래와 같이 표현됨을 보이시오.

$$\boldsymbol{e} = (e_1, e_2, \ldots, e_n)' = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}$$

여기서, $\boldsymbol{H} = \{h_{ij}\} = \boldsymbol{X(X'X)^{-1}X'}, \quad i, j = 1, 2, \ldots, n.$

(b) 잔차벡터의 기댓값 $E(\boldsymbol{e})$ 을 구하시오.

(c) $Var(e_i)$을 구하시오.

(d) 만약 $i \neq j$일 때, $Cov(e_i, e_j)$를 구하시오.

(e) 잔차들은 서로 독립인가? (TRUE or FALSE)

2. 모자행렬 $\boldsymbol{H}$의 대각원소 합이 p가 됨을 보여라. (힌트: $tr(\boldsymbol{AB}) = tr(\boldsymbol{BA})$)

3. 단순선형회귀에서 반응변수와 설명변수를 표준화하면 기울기의 추정값 $\widehat{\beta}_1$은 두 변수의 표본상관계수 r로 주어짐을 보여라.

4. 회귀분석의 분산분석표에 주어지는 F비에 대한 검정도 부분 F 검정에 해당된다. 이 경우 완전모형과 축소모형은 각각 어떻게 주어지나? 설명변수가 2개 있다고 가정하여라.

5. 수준 i에서의 예측값을 $\widehat{Y}_i$라 표기할 때, 아래 등식을 사용하여

$$(Y_{il} - \widehat{Y}_i) = (Y_{il} - \overline{Y}_i) + (\overline{Y}_i - \widehat{Y}_i), \quad i = 1, 2, \ldots, m;\ l = 1, 2, \ldots, n_i$$

다음의 제곱합 분할이 성립함을 증명하여라.

$$\sum_{i=1}^{m}\sum_{l=1}^{n_i}(Y_{il} - \widehat{Y}_i)^2 = \sum_{i=1}^{m}\sum_{l=1}^{n_i}(Y_{il} - \overline{Y}_i)^2 + \sum_{i=1}^{m} n_i(\overline{Y}_i - \widehat{Y}_i)^2$$

6. 자료 'dat1'에 포함된 설명변수 Y와 반응변수 $X_1, \ldots, X_4$는 다음과 같다.

```
> str( dat1)
'data.frame': 25 obs. of  6 variables:
 $ Y  : int  264 240 288 228 240 219 174 348 312 297 ...
 $ X1 : num  57.3 41.3 73.3 67.3 66.7 ...
 $ X2 : num  73.3 64.7 71.3 78 67.3 ...
 $ X3 : num  66.7 66 68.7 62 63.3 ...
 $ X4 : num  58 66.7 68.7 63.3 58.7 ...
```

(a) 완전모형을 적합하는 R 코드를 적어 보아라.

(b) 가설 $H_0 : \beta_2 = 0,\ \beta_3 = 2.5\beta_4$ 하에서 주어지는 축소모형을 적합하는 R 코드를 적어 보아라.

(c) 만약 적합된 완전모형 $fit1$과 부분모형 $fit2$의 결과를 사용하여 부분 F 검정을 실행하기 위해 다음의 코드를 실행하였다 (실행 결과의 일부는 가림처림함). 부분 F 검정에 사용되는 통계량의 값을 적으시오.

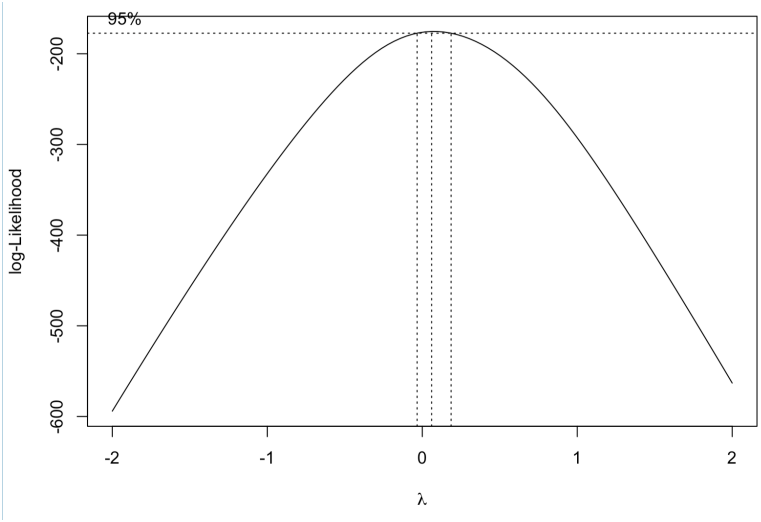
```
> anova( fit1 )
Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value    Pr(>F)
X1         1  21567   21567 142.663 1.476e-10 ***
X2         1  16261   16261 107.562 1.708e-09 ***
X3         1  38293   38293 253.302 8.032e-13 ***
X4         1   2342    2342  15.493 0.0008169 ***
Residuals 20   3023     151

> anova( fit2, fit1 )
      RSS
fit1 3136.4
fit2 3023.5
```

(d) 가설 $H_0 : \beta_2 = 0,\ \beta_3 = 2.5\beta_4$을 기각할 수 있는가? 그 이유를 간략히 설명해 보시오.

7. 아래는 자료에 포함된 반응변수의 λ 그림이다.



(a) λ 그림은 회귀분석의 어떤 가정과 관련이 있나?

(b) λ 그림을 참고하여 적절한 박스-칵스 변환모수 λ를 선택하라. 이 값은 어떤 변환에 해당되는가?

8. 자료 'dat2'에 반응변수 Y와 설명변수 x_1, x_2, x_3가 포함되어 있다. 자료에 포함된 설명변수들을 사용하여 아래와 같은 결과를 얻었다.

```
> R <- cor(dat2[2:4])
> eig <- eigen(R)
> diag(solve(R))
      x1       x2       x3
708.8429 564.3434 104.6060
> eig
eigen() decomposition
$values
[1] 2.0664726783 0.9328007024 0.0007266194

$vectors
           [,1]        [,2]       [,3]
[1,] -0.6946957 -0.05010563  0.7175565
[2,] -0.6294279 -0.44050902 -0.6401347
[3,] -0.3481645  0.89634883 -0.2744818
> summary( vif( lm( y ~ ., dat2) ) )

...

Variance Proportion:
   Eigenvalues Cond.Index           x1           x2           x3
1 2.0664726783   1.000000 3.294652e-04 0.0003397182 0.0005607672
2 0.9328007024   1.488403 3.796945e-06 0.0003686187 0.0082339577
3 0.0007266194  53.328743 9.996667e-01 0.9992916631 0.9912052751
```

(a) 설명변수를 모두 포함하는 모형을 적합하였을 때 다중공선성의 가능성과 몇 개의 선형종속관계가 존재하는지 판단해 보아라.

(b) 조건수(condition number)와 조건지수(condition index)를 구하라.

(c) 분산비(variance proportion) 값을 참고하여 다중공선성의 형태를 설명하여라.

9. 자료에 단순선형회귀모형을 적합한 후 모형 진단을 실행한 결과이다.

```
> dwtest(fit)

 Durbin-Watson test

data:  fit
DW = 2.3391, p-value = 0.9124
alternative hypothesis: true autocorrelation is greater than 0
```

(a) 더빈-왓슨 검정은 회귀분석의 어떤 가정과 연관되어 있나?

(b) 위 결과를 참고하여

i. 귀무가설과 대립가설을 적어라.

ii. 가설 검정 결과를 설명하여라.

10. 다음은 자료에 선형모형을 적합한 후 모형진단 및 영향력 측도를 계산한 결과이다.

```
> str(dat)
'data.frame': 10 obs. of  5 variables:
 $ y : int  790 1380 270 1190 590 1120 815 450 1290 420
 $ x1: int  78 39 109 20 70 58 53 68 15 96
 $ x2: int  133462 33000 120000 69727 112000 39106 95935 120000 20215 140000
 $ x3: int  1998 2000 1800 1999 2000 1998 1800 1800 1798 1800
 $ x4: int  1 1 0 1 0 1 1 0 1 0
> fit <- lm( y ~ ., dat)
> hatvalues(fit)
        1         2         3         4         5
0.7036548 0.3550714 0.5246524 0.4917191 0.5954326
        6         7         8         9        10
0.5249483 0.4817434 0.3858682 0.6203192 0.3165906
> rstudent(fit)
         1          2          3          4          5
 0.3751797  1.7958104 -0.8905142 -0.3407665 -0.6598631
         6          7          8          9         10
-1.0483361 -0.7927159 -0.3423444  0.2223592  2.2019925
> cooks.distance( fit )
         1          2          3          4          5
0.08071616 0.24574863 0.18261361 0.02729222 0.14448233
         6          7          8          9         10
0.23817214 0.12620463 0.01788553 0.01994855 0.25384332
```

(a) 이상치가 존재하는지 판단해 보아라. 어떤 통계량을 사용해서 판단하였나?

(b) 높은 지렛점이 존재하는지 판단해 보아라. 어떤 통계량을 사용해서 판단하였나?

(c) 쿡의 거리 통계량은 어떤 통계량에 영향력을 판단해 보기 위한 측도인가?