회귀분석1 2020년 1학기 기말고사 학번 _____ 성명 _____

배점: 1-25, 2-15, 3-50, 4-40, 5-15 : 합계 145점

1. (1) 반응변수 Y가 kg 단위를 가질 때 오차제곱합 SSE와 결정계수 R^2는 각각 어떠한 단위를 가지는가?

(2) 설명변수가 4개 있는 회귀모형 $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$ 에서 가설 $H_0 : \beta_1 = 0 , \beta_3 - \beta_4 = 0$에 대한 F 검정을 수행하고자 한다. 이 경우에 완전모형과 축소모형은 각각 어떻게 주어지는가?

(3) PRESS는 $\sum_{i=1}^{n} (y_i - \hat{y}_{i(i)})^2$ 으로 정의된다. 이 공식에서 $\hat{y}_{i(i)}$는 무엇인가?

(4) 분산안정화변환(variance stabilizing transformation)이란 무엇인가?

(5) 한 모집단에서의 평균을 추정하는 문제에 대한 통계모형은 다음과 같이 설정된다.

$$Y_i = \mu + \varepsilon_i , \quad i = 1, \cdots, n$$

Y_i의 모평균 μ의 최소제곱추정량은 Y_i의 표본평균 $\overline{Y}$로 주어짐을 보여라.

(6) 모형선택 문제에서 사용되는 간결함의 원칙(principle of parsimony)을 간결하게 설명하여라.

(7) (True/False) 회귀분석에서 설명변수들의 편제곱합을 모두 더하면 회귀제곱합(SSR)이 된다.

(8) (True/False) 잔차들은 서로 독립이고 동일한 분산을 가진다.

2. 반응변수 Y와 설명변수 X_1의 산점도에서 $X_1 = 54$ 근처에서 불연속점(discontinuity point)의 가능성이 나타나 이것을 구체적으로 조사하기 위해 다음의 모형을 적합하였다.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 Z_2 + \beta_3 X_a + \varepsilon$$

위의 모형에서 Z_2는 X_1이 54 이하이면 0, 54보다 크면 1의 값을 갖는 가변수이고, $X_a = (X_1 - 54) \cdot Z_2$이다. 위의 모형을 적합한 결과가 다음과 같을 때 각 질문에 답하여라.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)     30.5      3.640    8.38 0.000157
X1               2.83     0.361    7.84 0.000228
Z2               1.24     0.408    3.04 0.022803
Xa              -1.75     0.412   -4.25 0.005380
```

(1) X_1이 54 이하인 경우와 54 보다 큰 경우에 대해 각각 해당되는 모집단 회귀식을 구하여라.

(2) 위의 두 모집단 회귀식에서 각각 X_1=54를 대입한 후 그 차이를 계산하여라.

(3) (1)에서 구한 두 회귀식을 그림으로 나타내어라.

(4) 위의 적합된 결과를 이용하여 불연속점의 가능성에 대해 검정하여라.

(5) 위의 적합된 결과를 이용하여 두 부분의 기울기가 같다는 가설을 검정하여라.

3. 다음은 4개의 지역(region)에서 성장하는 어느 열대나무에 대한 분석이다.

관련변수: 반응변수-무게(weight, kg), 설명변수-지름(diameter, cm), 지역(region)

변수 region은 범주형변수(1, 2, 3, 4)이므로 가변수로 변환되어 사용된다.

```
tree <- read.table("D:/tree2.txt")
names(tree) <- c("region", "diameter", "weight")
head(tree)
```

```
  region diameter weight
1      1      7.2 10.404
2      1      8.2 18.161
3      1     10.3 25.778
4      1     10.1 20.511
5      1     10.7 21.870
6      1     13.3 47.186
```

 # Model 1

```
fit1 <- lm(weight ~ factor(region) - 1, data=tree)
summary(fit1)
```

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
factor(region)1   23.985      6.062   3.956 0.000779 ***
factor(region)2   27.243      6.062   4.494 0.000222 ***
factor(region)3   16.430      6.062   2.710 0.013472 *
factor(region)4   16.582      6.062   2.735 0.012752 *

Residual standard error: 14.85 on 20 degrees of freedom
Multiple R-squared:  0.717,     Adjusted R-squared:  0.6604
F-statistic: 12.67 on 4 and 20 DF,  p-value: 2.69e-05
```

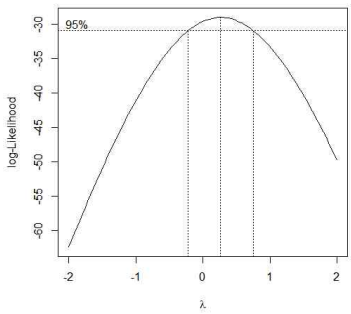
***** Model 1에 대한 질문**

(1) 추정된 회귀식은 무엇인가?

(2) 상수항이 제외된 이유는 무엇인가?

(3) 각 회귀계수 추정치의 의미는 무엇인가?

(4) 분석에 사용된 자료의 수는 몇 개인가?



[그림 1]

 # Model 2

```
fit2 <- lm(weight ~ factor(region), data=tree)
summary(fit2)
```

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       23.985      6.062   3.956 0.000779 ***
factor(region)2    3.258      8.573   0.380 0.707909
factor(region)3   -7.555      8.573  -0.881 0.388672
factor(region)4   -7.403      8.573  -0.863 0.398100

Residual standard error: 14.85 on 20 degrees of freedom
F-statistic: 0.8007 on 3 and 20 DF,  p-value: 0.508
```

***** Model 2에 대한 질문**

(5) 추정된 회귀식은 무엇인가?

(6) 지역1과 지역3에서의 평균 무게 차이에 대한 추정값은 얼마인가? 이 차이는 유의한가?

(7) 위의 모형에 변수 diameter가 설명변수로 포함되어 회귀계수 추정값 $\hat{\beta}$이 나왔다고 가정하자. 이때 반응변수 weight(kg)의 단위를 g으로, 설명변수 diameter(cm)의 단위를 mm로 변환하면 회귀계수 추정값 $\hat{\beta}$은 어떻게 변하는가?

(8) [그림 1]은 fit2에 대해 boxcox() 함수를 적용한 결과이다. 해석하여라.

 # Model 3

```
fit3 <- lm(log(weight) ~ log(diameter) + factor(region) + log(diameter):factor(region), data=tree)
summary(fit3)
```

```
Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  -1.88763    0.72793  -2.593   0.0196 *
log(diameter)                 2.17653    0.31807   6.843 3.95e-06 ***
factor(region)2              -0.32517    0.81687  -0.398   0.6958
factor(region)3               0.10403    0.82410   0.126   0.9011
factor(region)4              -0.18701    0.85696  -0.218   0.8300
log(diameter):factor(region)2  0.08167    0.35593   0.229   0.8214
log(diameter):factor(region)3 -0.16326    0.36374  -0.449   0.6596
log(diameter):factor(region)4 -0.01637    0.37839  -0.043   0.9660

Residual standard error: 0.1529 on 16 degrees of freedom
Multiple R-squared:  0.9709,    Adjusted R-squared:  0.9581
F-statistic: 76.19 on 7 and 16 DF,  p-value: 4.308e-11
```

***** Model 3에 대한 질문**

(9) 지역2에 대한 추정회귀식을 적어라.

(10) 회귀계수 추정값 0.08167의 의미를 설명하여라.

(11) 지역1과 지역2에 대한 회귀식에서 절편의 차이와 기울기의 차이를 각각 검정하여라.

 # Model 4

```
fit4 <- lm(log(weight) ~ log(diameter) + factor(region), data=tree)
summary(fit4)
```

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    -1.83481    0.21905  -8.376 8.41e-08 ***
log(diameter)   2.15337    0.09250  23.280 1.98e-15 ***
factor(region)2 -0.13821    0.08368  -1.652  0.11504
factor(region)3 -0.25147    0.08441  -2.979  0.00771 **
factor(region)4 -0.22498    0.08414  -2.674  0.01501 *

Residual standard error: 0.1449 on 19 degrees of freedom
Multiple R-squared:  0.9689,    Adjusted R-squared:  0.9624
F-statistic: 148.1 on 4 and 19 DF,  p-value: 4.852e-14
```

***** Model 4에 대한 질문**

(12) 부분 F 검정을 이용하여 Model 3에서 2차 교호작용항들이 필요한지를 검정하여라.

(13) 'log(diameter)'의 계수 β_1에 대한 95% 신뢰구간을 구하여라.

(14) 'log(diameter)'의 계수 β_1에 대한 가설 $H_0 : \beta_1 = 1.9$, $H_1 : \beta_1 \neq 1.9$ 을 검정하여라.

(15) $\hat{\beta}_1$의 분산인 $Var(\hat{\beta}_1)$의 추정값은 얼마인가?

(16) 위의 분석에서 오차분산 σ^2의 추정값은 얼마인가?

4. 다음은 교재 3장과 6장의 연습문제에서 사용되었던 프로 야구선수들의 연봉과 관련한 7개의 설명변수들에 대해 회귀분석을 한 결과이다. 각 물음에 답하여라.
(참고: 변환된 연봉을 사용함.)
설정된 모형 : $newY = \beta_0 + \beta_1 X1 + \beta_2 X2 + \beta_3 X3 + \beta_4 X4 + \beta_5 X5 + \beta_6 X6 + \beta_7 X7 + \varepsilon$

[출력물 1]

```
KBO1$newY = log10(KBO1$Y) *1000
summaryf(regsubsets(newY ~ X1 + X2 + X3 + X4 + X5 + X6 + X7, KBO1, nbest=3))
```

```
         X1 X2 X3 X4 X5 X6 X7     rss     rsq   adjr2       cp     bic    PRESS
1  ( 1 )                  *    6403060 0.73616 0.73451 194.9438 -205.67  6672706
1  ( 2 )  *                    8974603 0.63019 0.62788 336.6902 -150.98  9206715
1  ( 3 )              *        9758671 0.59788 0.59537 379.9089 -137.41  9982347
2  ( 1 )  *               *    3219891 0.86732 0.86565  21.4840 -311.95  3444851
2  ( 2 )              *   *    3356064 0.86171 0.85997  28.9900 -305.24  3578066
2  ( 3 )        *         *    3428497 0.85873 0.85695  32.9826 -301.78  3641716
3  ( 1 )  *               *  * 2985239 0.87699 0.87465  10.5497 -319.12  3215247
3  ( 2 )  *            *  *    3060848 0.87387 0.87148  14.7174 -315.07  3317406
3  ( 3 )  *      *        *    3109292 0.87188 0.86945  17.3877 -312.52  3366655
4  ( 1 )  *            *  *  * 2826972 0.88351 0.88054   3.8259 -322.86  3087241
4  ( 2 )  *      *        *  * 2901351 0.88045 0.87740   7.9257 -318.65  3170830
4  ( 3 )  *  *             *  * 2926101 0.87943 0.87636   9.2900 -317.27  3185035
5  ( 1 )  *      *     *  *  * 2798100 0.88470 0.88101   4.2344 -319.43  3104240
5  ( 2 )  *  *           *  *  * 2804915 0.88442 0.88072   4.6101 -319.04  3095588
5  ( 3 )  *         *  *  *  * 2817569 0.88390 0.88018   5.3076 -318.31  3122884
6  ( 1 )  *      *  *  *  *  * 2794684 0.88484 0.88038   6.0461 -314.54  3145122
6  ( 2 )  *  *  *     *  *  * 2797635 0.88472 0.88026   6.2088 -314.37  3159586
6  ( 3 )  *  *     *  *  *  * 2798489 0.88469 0.88022   6.2559 -314.32  3134426
7  ( 1 )  *  *  *  *  *  *  * 2793847 0.88488 0.87964   8.0000 -309.50  3199826
```

(1) [출력물 1]을 참조하여 모형선택 기준으로 '수정된 결정계수', '맬로우즈 C_p', 'BIC', 'PRESS를 사용할 때 선택되는 모형을 각각 적어라.

[출력물 2] : Model 1

```
fit1 <- lm(newY ~ X1 + X2 + X3 + X4 + X5 + X6 + X7, data=KBO1)
summary(fit1)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3402.81696   30.66093 110.982  < 2e-16 ***
X1             2.48695    1.15108   2.161 0.032276 *
X2             1.08313    5.04473   0.215 0.830281
X3             0.98934    1.95585   0.506 0.613694
X4            -0.87068    1.90541  -0.457 0.648350
X5             3.66287    1.55127   2.361 0.019467 *
X6             0.43701    0.04238  10.311  < 2e-16 ***
X7            12.45796    3.66209   3.402 0.000853 ***

Residual standard error: 134.7 on 154 degrees of freedom
Multiple R-squared:  0.8849,    Adjusted R-squared:  0.8796
F-statistic: 169.1 on 7 and 154 DF,  p-value: < 2.2e-16
```

```
summary(vif(fit1))
```

```
VIF:
   X1    X2    X3    X4    X5    X6    X7
25.72  6.72 20.30 18.48  7.93  2.89  2.05
```

```
Variance Proportion:
  Eigenvalues Cond.Index         X1         X2         X3        X4        X5         X6         X7
1    4.959201     1.0000 0.00143724 4.3803e-03 0.00182526 0.0019161 0.0044911 0.0070282 0.00334790
2    1.233973     2.0047 0.00065091 2.2875e-03 0.00091503 0.0013423 0.0020794 0.0798404 0.23792689
3    0.397326     3.5329 0.00751518 2.2010e-01 0.00863472 0.0189235 0.0146731 0.0451210 0.03509586
4    0.230743     4.6360 0.00226230 2.3454e-02 0.00018960 0.0057245 0.0179209 0.8406168 0.70188835
5    0.102953     6.9404 0.07310091 6.0026e-05 0.00805499 0.0409848 0.8679144 0.0051251 0.00786510
6    0.055133     9.4842 0.02923598 4.5884e-01 0.42105002 0.3085543 0.0015895 0.0033703 0.01352608
7    0.020671    15.4889 0.88579749 2.9088e-01 0.55933038 0.6225545 0.0913315 0.0188981 0.00034982
```

(2) [출력물 2]에서 다중공선성 여부를 판단하여라.

(3) 만약 변수들에 선형종속이 1개 있는 경우 어느 변수들 간에 있을 가능성이 높은가?

[출력물 3] : Model 2

```
fit2 <- lm(newY ~ X1 + X5 + X6 + X7, data=KBO1)
summary(fit2)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.397e+03  3.013e+01 112.732  < 2e-16 ***
X1          2.520e+00  5.627e-01   4.478 1.44e-05 ***
X5          3.937e+00  1.328e+00   2.965  0.00350 **
X6          4.349e-01  4.203e-02  10.349  < 2e-16 ***
X7          1.302e+01  3.614e+00   3.604  0.00042 ***

Residual standard error: 134.2 on 157 degrees of freedom
Multiple R-squared:  0.8835,    Adjusted R-squared:  0.8805
F-statistic: 297.7 on 4 and 157 DF,  p-value: < 2.2e-16

anova(fit2)
Analysis of Variance Table

Response: newY
           Df   Sum Sq  Mean Sq F value    Pr(>F)
X1          1 15293718 15293718 849.359 < 2.2e-16 ***
X5          1   365604   365604  20.304 1.284e-05 ***
X6          1  5548152  5548152 308.125 < 2.2e-16 ***
X7          1   233876   233876  12.989 0.0004203 ***
Residuals 157  2826972    18006

drop1(fit2, test="F")
Single term deletions

Model:
newY ~ X1 + X5 + X6 + X7
       Df Sum of Sq     RSS    AIC  F value    Pr(>F)
<none>              2826972 1592.3
X1      1    361056 3188028 1609.8  20.0518 1.443e-05 ***
X5      1    158267 2985239 1599.1   8.7896 0.0035025 **
X6      1   1928347 4755320 1674.5 107.0936 < 2.2e-16 ***
X7      1    233876 3060848 1603.2  12.9886 0.0004203 ***

lmtest::dwtest(fit2)

        Durbin-Watson test

data:  fit2
DW = 1.8352, p-value = 0.1253
alternative hypothesis: true autocorrelation is greater than 0
```

(4) 부분 F 검정을 이용하여 Model 1, Model 2 중에서 어느 모형이 좋은지를 판정하여라.

(5) 더빈-왓슨 검정결과를 해석하여라.

(6) 'Model 2'에서 $H_0 : \beta_1 = \beta_5 = \beta_6 = \beta_7 = 0$에 대한 검정을 실시하여라.

(7) 순차제곱합의 의미를 간략하게 설명하여라. 그리고 변수 'X5'의 순차제곱합은 얼마인가?

(8) 편제곱합의 의미를 간략하게 설명하여라. 그리고 변수 'X5'의 편제곱합은 얼마인가?

(9) 위에서 변수 'X7'의 순차제곱합과 편제곱합이 같은 이유는 무엇인가?

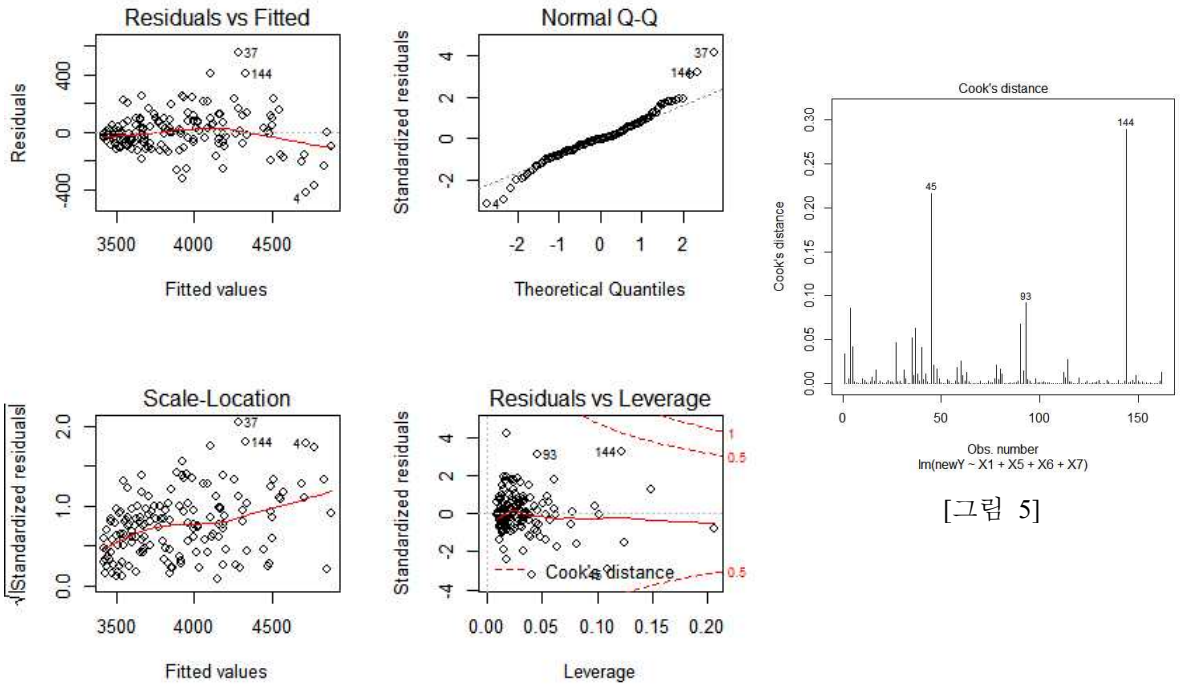
(10) 'Model 2'의 4개 설명변수들 중에서 반응변수에 제일 중요하게 영향을 주는 변수는 무엇인가? 그 이유는?

(11) [출력물1]에서 'Model 2'의 press값은 얼마인가? 'Model 2'의 오차제곱합(SSE)와 크기를 비교하여라. 큰 차이가 나는 경우 그것이 의미하는 것은 무엇인가?

(12) 'Model 2'에 대한 $R^2_{Prediction}$의 값을 구하고, R^2와 비교하여라.

5. 다음은 앞의 모형에 대해 추가적인 분석을 실시한 결과이다.
(답을 적을 때 그림 번호를 반드시 표시할 것.)

plot(fit2) (그림 번호: 좌상-1, 우상-2, 좌하-3, 우하-4)



[그림 5]

(1) 오차항에 대한 가정 중 등분산성과 정규성이 성립하는지 설명하여라.

(2) "특이점"은 몇 개 정도 있는지를 판단하여라.

(3) 큰 지렛값에 대한 기준값이 0.12이라면 "높은 지렛점"은 몇 개 정도 있는가?

(4) 기준값 0.1를 사용할 때 Cook의 거리 측도에서 나타나는 영향력 관측치는 몇 개인가? Cook의 거리 측도는 무엇에 미치는 영향력 정도를 나타내는가?